AI Weather Forecasts for Health

The case for a decision-first approach

Full Report · September 2026

THE UNIVERSITY OF CHICAGO
INSTITUTE FOR CLIMATE & SUSTAINABLE GROWTH

Supported by The Rockefeller Foundation

Report drafting process and AI use disclaimer

This work was built by convening an expert group as well as desk review. The report draws on multiple interviews with 30 experts conducted between March 2026 and July 2026. Draft findings were prepared and approximately 40 participants spanning meteorology, AI weather modeling, public health, humanitarian response, technology policy, and philanthropic and multilateral funding met in London over two days in June 2026. What follows reflects that discussion in response to the draft findings. Inputs from participant interviews and the expert convening were sometimes processed using Anthropic Claude models for clarity before being used as sources for drafting. AI summary notes through Zoom were used to supplement the author team's notes from interviews. Where possible, full transcripts with text clarified using Zoom or Claude AI tools were used instead of AI summaries.

This report was written, dictated, or drafted by the author team. During the drafting of the report, Claude models were used to copyedit or suggest minimal polishing of text to improve readability and reduce copyediting errors. After using this tool, the authors reviewed and edited all content and take full responsibility for the final publication content and any errors contained herein.

Acknowledgements

This report was written by Amir Jina, Harris School of Public Policy, University of Chicago; Kristina Costa, Institute for Climate and Sustainable Growth, University of Chicago; Madyson Miller, Data Science Institute, University of Chicago; and Juli Trtanj, Trtanj Consulting. Design by Ryan Reynolds, Philament Design.

The author team would like to express their thanks to those who reviewed and/or co-wrote sections of the text: Pedram Hassanzadeh, Katie Kowal, and Kate Burrows, University of Chicago; Greg Kuzmak, The Rockefeller Foundation; Hunter Jones, World Meteorological Organization; Shona Kamps, WHO-WMO Joint Office for Climate and Health; Helen Caughey, Abigail Odat, Nyree Pinder, Met Office (UK); Bijal Brahmbhatt, Chintan Shinde, Siraz Hirani, Gujarat Mahila Housing Sewa Trust; Felipe Gonzalez Colon, Wellcome Trust; Yang Liu, PATH; Titi Akinsanmi, Demistef.ai; Roxy Mathew Koll, Indian Institute of Tropical Meteorology; Caradee Wright, South African Medical Research Council; Imara Salas, AIM for Scale; Jeffrey Shrader, Columbia University; Kyle Joseph; Chen Hall, University of Maryland.

A special thanks to collaborating experts who donated their time to share insight and provide expertise, including: Michael Greenstone and Michael Kremer, University of Chicago; Madeleine Thomson, Wellcome Trust; Bilal Mateen, PATH; Karthik Kashinath and Niall Robinson, NVIDIA; Joy Shumake-Guillemot, World Meteorological Organization; Jenna Allard, Yale University; Julie Arrighi, Red Cross Red Crescent Climate Centre & American Red Cross; Timothy A. Bouley, Stanford University; John Bosco Isunju, Makerere University School of Public Health; Matthew Chantry, ECMWF; Tufa Dinku, Stockholm Resilience Centre; Joyce Kimutai, Imperial College London; Rachel Lowe, Barcelona Supercomputing Center; Gabriel Moldovan, ECMWF; Virginia Murray, Global Hazard Facility (UNDRR); Ousmane Ndiaye, African Centre of Meteorological Applications for Development (ACMAD); Shuchita Rawal, Civic Data Lab; James Smallcombe, Heat and Health Research Centre, University of Sydney; Naveen Rao and Emma Dahill, The Rockefeller Foundation; Rich Turner, Cambridge University.

Responsibility for any remaining errors rests with the lead author.

This report was made possible with support from The Rockefeller Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of The Rockefeller Foundation.

Table of Contents

Acronyms

ACMAD	African Centre of Meteorological Applications for Development
AIFS	Artificial Intelligence Forecasting System (ECMWF)
AIM for Scale	Agricultural Innovation Mechanism for Scale
AIWP	Artificial Intelligence Weather Prediction
API	Application Programming Interface
CHI	Climate-Health Intelligence
CH-MIP	Climate-Health Model Intercomparison Project
CHW	Community Health Worker
COPD	Chronic Obstructive Pulmonary Disease
CPU	Central Processing Unit
CSfH	Climate Services for Health
DALY	Disability Adjusted Life Year
DHIS2	District Health Information Software 2
D-MOSS	Dengue Forecasting Model Satellite-based System
DPI	Digital Public Infrastructure
ECMWF	European Centre for Medium-Range Weather Forecasts
ENACTS	Enhancing National Climate Services
ENSO	El Niño-Southern Oscillation
ERA5	ECMWF Reanalysis v5
EW4All	Early Warnings for All
FAIR	Findable, Accessible, Interoperable, Reusable
FEWS NET	Famine Early Warning Systems Network
GBD	Global Burden of Disease
GPU	Graphics Processing Unit
HL7 FHIR	Health Level Seven Fast Healthcare Interoperability Resources
HNAP	Health National Adaptation Plan
IHR	International Health Regulations
IITM	Indian Institute of Tropical Meteorology
IPCC	Intergovernmental Panel on Climate Change
LMIC	Low- and Middle-Income Countries
MDB	Multilateral Development Banks
NAP	National Adaptation Plan
NMHS	National Meteorological and Hydrological Service
NWP	Numerical Weather Prediction
PM2.5	Particulate Matter under 2.5 micrometers
RCT	Randomized Control Trial
REV	Relative Economic Value
S2S	Subseasonal-to-Seasonal Forecasting
SAMRC	South African Medical Research Council
SAWS	South African Weather Service
SOFF	Systematic Observations Financing Facility
UTCI	Universal Thermal Climate Index
VBD	Vector Borne Disease
VIMC	Vaccine Impact Modeling Consortium
WBGT	Wet-bulb Globe Temperature
WHO	World Health Organization
WIS 2.0	WMO Information System 2.0
WISER	Weather and Climate Information Services for Africa
WMO	World Meteorological Organization

Foreword

Our health, both physical and mental, is one of the principal ways we experience the impacts of weather and climate. As heatwaves, drought and flooding intensify, so does their toll on health. Yet despite these growing risks, weather and climate information remains insufficiently integrated into routine health decision-making. In many countries, health leaders making decisions about staffing, supplies, preparedness, or public health interventions still do so without access to tailored climate and weather intelligence. As our climate continues to change the frequency and severity of many health-related hazards, closing this gap becomes ever more urgent.

For decades, the challenge has not only been scientific but operational. Developing and maintaining forecasting systems capable of supporting specific health decisions has required significant investments in infrastructure, expertise, observations, and computing power. Many National Meteorological and Hydrological Services, particularly in developing countries, have relied on adapting forecasting systems developed elsewhere, often with limited ability to tailor them to local contexts and health priorities. As a result, climate and weather information has too often remained peripheral to health planning rather than becoming an integral part of it.

Artificial intelligence is beginning to change this equation. AI-enabled prediction systems offer the prospect of producing forecasts more quickly, more efficiently, and at a scale that was previously unimaginable. They are lowering technical barriers and creating new opportunities for countries to develop forecasting services tailored to their own needs. Importantly, they open the possibility of designing prediction systems not only around meteorological variables, but around the health decisions they are intended to inform.

This potential is significant. It offers an opportunity to accelerate what WMO refers to as climate-health intelligence: the systematic integration of weather, climate, environmental, and health information into decision-making processes that protect lives and strengthen resilience. By helping reveal relationships between climate risks and health outcomes, AI may enable earlier interventions, more targeted preparedness measures, and more effective use of limited health resources.

However, technological capability alone will not translate into better outcomes. The deployment of AI in climate and health forecasting raises equally important questions of governance, responsibility, and trust. Health systems do not act on information simply because it is available; they act on information they trust. Trust arises not only from scientific performance, but from clear institutional mandates, transparent methodologies, and well-defined accountability across the prediction-to-action chain. In most countries, these responsibilities are shared across National Meteorological and Hydrological Services, Ministries of Health, public health authorities, research institutions, and emergency management agencies. Strengthening the governance arrangements that connect these actors will be as important as advancing the technology itself.

This report recognizes that the future of climate-health forecasting depends on both innovation and stewardship. It examines how emerging AI capabilities can support more decision-relevant forecasting while highlighting the institutional foundations required for their responsible use. Success should not be measured solely by improvements in forecast accuracy, but by whether decision-makers receive information they can act upon confidently and effectively.

AI has made forecasting systems faster to build, easier to customize, and more accessible than ever before. Yet it has not answered the fundamental questions of governance, accountability, and public trust that accompany their adoption. Those questions must be addressed now, while the next generation of climate-health services is being shaped.

Véronique Bouchet
Senior Director, Department of Science, Services and Capacity Development,
World Meteorological Organization

Summary for Decision-makers

Health systems make decisions that the weather decides for them. When to open cooling centers and how many people to staff them. When to run a malaria control spraying round. Whether a district hospital should staff for a surge next week or next month. Each of these decisions has an owner, a budget, and a deadline, and each has a point in time when better information about the coming weather would change what decision is made—and, potentially, how many lives are saved.

Over three decades, a community of researchers and practitioners has established how heat, rainfall, and drought translate into illness. It built the first heat action plans and first health-triggered warning systems. It assembled the national climate records in countries whose station networks were too sparse to support them. And, through direct collaboration with the people who make these decisions, it worked out many ways that a service might work in practice. The expertise and relationships exist. What never existed is a forecast system capable of being built to truly serve these decisions.

The reason is rooted in a much older history of weather forecasting. Seventy years ago, advances in computing and physics enabled a first revolution in weather forecasting. These physics-based models came with several key constraints. Running them required massive upfront investments in heavy-duty supercomputers, costing on the order of a hundred million dollars, alongside robust systems for local data collection and aggregation. Building forecasts around specific users' needs took months or years. Adapting forecasts to local conditions was expensive enough that only a few meteorological services globally could do it. It is unsurprising, then, that modern weather forecasts have long been most accurate in the wealthiest parts of the planet—places like Europe and the United States, where national meteorological agencies have been able to invest for decades in iterative modeling and computing improvements and where consistent, ongoing support has resulted in the densest networks of monitoring stations.

We are currently experiencing a second revolution in weather forecasting. AI weather models can now match or beat the best physics-based forecasts, at a fraction of the cost per forecast and relying on no more computing power than a laptop and they have lifted these old constraints. That means we can now design a climate service around the health decision it needs to support. This requires "demand articulation"—defining what has to be decided, who owns that decision, what authority and resources they hold, and what would change with a better forecast. The forecast, evaluation, and delivery can then be built around those needs. This approach was available only to the wealthiest few weather services under the old constraints. It is now available to anyone. AI weather models and demand articulation together can turn three decades of accumulated methods into services operating at scale.

This second revolution is arriving at a moment when it is urgently needed. Climate change is upending historic weather patterns, increasing the severity of disaster events like downpours, floods, and droughts, and contributing to the spread of diseases like malaria. Many of the countries that are facing the most severe effects of climate change are the very same places in the tropics that have historically been left behind by the physics-based weather forecasting system.

The AI weather forecasting revolution is here. The only open question is whether policymakers, philanthropists, researchers, and technologists will harness and direct that revolution in a way that truly addresses the needs of those who can benefit most.

Key Findings

The following findings are drawn from literature review, dozens of expert interviews, and the results of an intensive, two-day-long convening in June 2026 that brought together researchers and practitioners from meteorology, public health, climate, technology, and the social sciences.

1

Steering the AI-driven second weather forecasting revolution toward health is a decision, and it is most effectively made now.

Left to its own incentives, AI weather forecasting adds skill fastest where data are densest, for the uses existing forecasts already serve well. The populations that the first forecasting revolution never truly reached may be left behind again. Nothing about the technology itself makes that inevitable. What determines it is a set of human choices about what data these models learn from, how they are evaluated, and whose decisions they are built to serve. Policymakers, researchers, and technologists should be deliberate about the choices they make now, because they get harder to reopen as benchmarks, standards, and procurement patterns become fixed around them. *(Chapter 1)*

2

AI weather forecasts remove the constraints that forced climate information for health to start from what could be produced instead of what users needed.

Running a leading forecast used to require a machine costing on the order of a hundred million dollars. It now takes minutes on a single processor. Rebuilding a model around one user's question took months, and now takes days. Adapting a forecast to local conditions is within reach of a modestly resourced weather service. None of that was true five years ago. "Supply-driven" climate services—the classic approach of producing a forecast and pushing it downstream to end users, rather than building it around those users' decisions—are now a choice, not a necessity. The time is now for weather services, health services, and funders to take advantage of what is newly possible. *(Chapters 1, 3)*

3

Last-mile problems in climate services are real, and a distinct first-mile problem has been folded in with them.

81% of national meteorological services report providing climate services for health, but just 23% of health ministries run a surveillance system that uses meteorological information. That gap is usually read as a dissemination problem, and some of it is. But a forecast built around what could be produced will often not fit the decision someone actually has to make, and low uptake is simply the first place that becomes visible. These "first-mile" problems can be addressed by decision owners specifying what a forecast must do before it is built. *(Chapter 2)*

4

The biology of the health impact and the logistics of the health system decide how much warning a decision needs, and which forecast can support it.

Some decisions need two to four weeks, enough for a health system to stage its response to a heat episode instead of reacting to it. Few forecasts exist at that lead time, so nobody asks for them. Other decisions need no forecast at all. Where a long biological lag exists between the weather and the illness, the lag supplies the warning time, and watching current conditions does more work than a forecast of what is coming. Investment has followed the forecast rather than the lead time the decision requires. Funders and service providers should start from the decision and its lead time, then invest in forecasts or monitoring accordingly. *(Chapter 2)*

5

Where forecast quality was the barrier, AI weather forecasting is removing it, and that is a substantial gain.

Forecasts have been too coarse to tell one district from another, too late to act on, or built around a weather variable that does not track what makes people ill. Improving climate services needs more, because other parts of the chain between a forecast and a healthier person are still missing. Poor forecasts gave funders a reason to neglect them. Health decisions were never specified, but health ministries can specify them. Forecasts were judged on weather accuracy and not health outcomes, but meteorological services can change what they evaluate against. Nobody collected the data to check if a warning helped, but funders can build that evidence. Institutions that need to work together were not required to speak to each other. While the forecast was the binding constraint, fixing any one of these alone changed little. Those difficulties remain. The return on addressing them has risen. *(Chapter 3)*

6

How models are evaluated is not only an issue of measurement. It decides which models are selected and developed next.

Comparing models against each other, against a reference forecast, and against what actually happened is called benchmarking, and developers optimize whatever the benchmark rewards. Today's AI weather benchmarks reward performance in data-rich regions by evaluating against the global atmosphere. They tell a health service little about whether a model supports decisions. A dengue alert should test models on dengue cases, and where no case record exists to run that test, it should test whether the forecast gets the weather threshold right. For a heat plan, that means the days it should activate, in the districts it covers. The same weather model can score well on one test and badly on the other. A better benchmark serves whoever specified the decision, and choosing whose decisions to specify is a separate and equally consequential step. *(Chapter 4)*

7

AI weather forecasting benefits will be greatest where they exist within a working system. This needs nine capacities with mandated owners. In most countries, several of them have none.

Demand articulation; weather data stewardship; health data stewardship; forecast creation; benchmarking; operational production; translation; delivery; and monitoring, evaluation, and feedback. Demand articulation comes first and is where the health sector drives the process: the decisions everything else is built around are stated by people who have to make them. Which institution should hold each capacity varies too much by country for a general recommendation. Cheaper AI forecasts will increase the number of warnings a health system has available. Who acts on a warning, and who decides not to, is usually unassigned. Every country can name a holder for each of the nine, and name who owns the action a warning triggers. *(Chapter 6)*

8

Better forecasts do not by themselves produce action. Confidence in a forecast's reliability does, together with accountability for errors.

Someone who will be blamed for a false alarm and not credited for a quiet season is unlikely to disseminate a forecast, and added accuracy does not change that calculation. Three things change this. The ability to analyze scenarios that let a decision-maker evaluate response options against the full probability distribution, not just a single number. Plain statements from the meteorological services of what a forecast can and cannot tell you, so that health users become critical readers of model output rather than recipients of it. And accountability for forecast errors settled in advance. That means a duty of care, not liability for outcomes. Met services and health ministries can settle accountability between them before the first warning goes out, so waiting to disseminate stops being the safest option. *(Chapters 4, 6)*

9

Equity considerations enter at the beginning of decision-driven AI forecast production. They are not a downstream benefit of having produced the right forecast.

Where an investment builds a forecast, a dataset, a benchmark, a research finding, or an institution, it has to state where decision-making, budgets, authorship, and long-term control will sit, and place them closer to the people carrying the greatest unmet burden. A model trained on African data, published by a European laboratory, and licensed back to African services satisfies every technical description of local relevance, and yet does not lead to any systematic change in forecast access or in local capacity. Without that specification, AI weather forecasting improves global forecasting capability and leaves the infrastructure concentrated exactly where it already is. Funders can require this in every proposal they approve. *(Chapters 6, 7)*

10

The AI policies governing how these forecasts and the data behind them can be used are being set now. Health is often written into them in name only, and changing this needs immediate effort.

Absent upfront decision-making, these policies risk name-checking areas like health without actually determining which agencies are responsible for taking particular actions. National climate adaptation plans provide a cautionary tale: Every one of 59 such plans reviewed in 2025 identifies health as a sector that is vulnerable to climate change, but fewer than half even identify a lead agency responsible for managing the issue. It's no wonder that just 0.2% of international climate adaptation finance has health as its primary focus. To prevent this, policymakers working on AI agendas should avoid naming health without any attached action or responsibility. The investment side of the same problem is set out below. *(Chapters 6, 7)*

Where investment is needed

Fully harnessing the potential of the AI revolution in weather forecasting to improve climate services for health means taking democratization and decentralization seriously from the outset. Actors across the growing AI weather for health ecosystem—from researchers to policymakers to technologists—should deliberately identify ways to situate decision-making, budgets, ownership, and long-term control of AI weather tools closer to the countries and communities where the need is greatest.

Investment needs and opportunities fall into four types, in sequence: Discovery, Evidence, Scaling, and Sustainability. Different actors are best positioned to take action at each stage. Evidence sits before Scaling deliberately, because a scaling funder cannot move without it. The right investments, at the right time, can push the AI weather for health field toward meeting the need where it is greatest. This illustrative list of examples is detailed in Chapter 7.

- **Discovery: Near-term catalytic and philanthropic investment.** Early investment by philanthropic funders will be essential to put AI weather forecasting on the right path for addressing the real-world needs of vulnerable populations. Example projects and initiatives include establishing a gold-standard reference dataset at the intersection of weather, climate, and health, so that future generations of AI weather models can be better trained; investing in critical operational tools like bias-correction and model calibration that frontier AI research will not focus on, as well as best practices for policy and governance; and building a consortium of researchers to collaborate on frontier discoveries, like subseasonal forecasts that don't exist today.
- **Evidence: Near- and medium-term philanthropic and research investment**. Philanthropic funders should coordinate investments to build a robust, transparent, and reliable evidence base for what works in AI weather for health. This should include research to answer the foundational questions, including whether weather warnings have measurably improved health outcomes, and what lead times key health decisions actually need. It should also include building support for evaluation of decision-making and health outcomes into grants for climate services and AI weather for health projects and pilots. One approach could be to establish a match-making facility for researchers and climate services implementers, which could make it easier for public agencies and NGOs to access a cadre of qualified researchers at the outset of new efforts to utilize AI weather for health.
- **Scaling: Medium- and long-term multilateral and private investment**. As researchers and agencies establish what works in the AI weather for health domain, multilateral development banks and private capital can take up the baton to enable replication and scaling of proven approaches and interventions. Establishing a climate-health scaling mechanism would bring innovations with strong evidence of health benefits to scale—an analog in agriculture has made substantial progress to accelerate the adoption of effective agriculture innovations among smallholder farmers. Alongside such an initiative, multilateral institutions can assist countries in effectively incorporating AI weather warning systems into existing heat action plans, later expanding to other hazards.
- **Sustainability: Long-term public- and private-sector investment**. In the long run, AI weather for health services will need to be accounted for in public-sector agency budgets, from training current and future meteorologists and health workers, to maintaining data sets and infrastructure. Based on the outcomes of research into the real economic value of AI forecasts for improving health, private-sector companies like hospitals and health insurers may be appropriate funders for certain kinds of sustainability activities.

The AI weather revolution is underway, and it will reach into public health and climate services regardless of whether those fields pay attention. This report calls on experts, practitioners, funders and the communities living on the front lines of climate change-driven weather and health impacts to engage now in shaping AI weather forecasts for health, so that they are judged on more than how well they reproduce the atmosphere. Without such engagement, AI weather forecasts will continue to answer questions the health sector has not asked.

CHAPTER 1

AI Weather Forecasts for Health: Why Now?

AT A GLANCE

- Weather and climate shape health on timescales from hours to seasons, and that spread of lags determines which forecasts can inform which health-related decisions.
- Forecasts should be driven by who needs information for a decision, what they need, and when they need it. Instead they are built around what forecasters can produce, with accuracy as the goal. This puts the burden on the "last mile" of dissemination, when the real problem is in the "first mile" where the decision was never specified.
- Historical experience and knowledge of past weather are no longer a reliable guide for action: variability is rising with climate change, and the weather events with the largest health consequences increasingly have no local precedent. Forecasts fill that gap.
- Three barriers made decision-driven forecasting infeasible: cost, iteration time, and local tailoring. AI relaxes all three at once, putting advanced forecasting within reach of a modestly resourced national service, and making the status quo a choice rather than a necessity.
- Left to its own incentives, AI weather improves fastest where data are densest, which is not where the health need is greatest. What determines the path is not the models themselves but the human decisions about how they are built, what data they use, how they are benchmarked, and who they are designed to serve.
- AI weather forecasting can bring large benefits, but the value is conditional. It removes the barriers that once made services that were driven by what could be forecast and not what was needed for decisions the only option. But it does not build the connections, the data, the institutions, or the governance that turn a forecast into a health outcome.

Climate and health

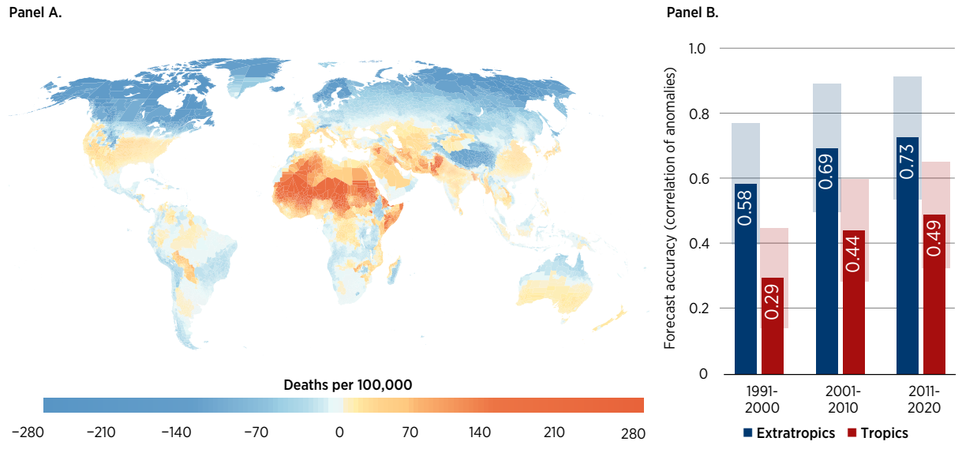
Weather and climate shape human health through heat, floods and storms, drought, food insecurity, poor air quality, and the shifting range of vector- and water-borne disease [1]. The scale is large and not fully understood: the IPCC puts the climate-sensitive disease burden at roughly 39.5 million deaths and 1.5 billion disability-adjusted life years in 2019, with cardiovascular disease the largest single category. Extreme heat alone accounts for around 489,000 deaths a year, and the true figure may be many times higher, because heat is seldom recorded as a cause of death and is often reported months after the season that caused it [2]. The direction of travel is clearer than the level: heat-related deaths rose 68% between 2000-2004 and 2017-2021, and the area with a climate suitable for dengue and malaria transmission has widened measurably over the past seventy years. Under future climate change, much of that burden concentrates in the tropics (Fig. 1.1), which is also where forecasting has historically been weakest [3].

What makes climate and health a forecasting problem and not only a planning one is that these outcomes occur on different timescales in response to weather or climate events. Heat kills within hours to days of exposure, which puts the useful decision inside a short-range window. Cholera and other diarrheal diseases follow flooding by days to weeks, long enough to pre-position supplies if the flood itself is anticipated. Vector-borne disease carries a longer biological lag again because mosquito development and pathogen incubation take weeks to months, so conditions observed today already imply cases some way ahead, and contemporary monitoring can do work no forecast can. Malnutrition tracks the growing and lean seasons. That spread of lags, from hours to seasons, is what determines which forecast could inform which decision [5].

> "This is not a data problem. It is a design problem. It is a trust problem. In some cases, it is a governance problem, and it is at its core an equity problem."
>
> **Naveen Rao**
> **The Rockefeller Foundation**

Climate services for health (CSfH) are the tools and information that help public health professionals anticipate, prepare for, and respond to the effects of

Figure 1.1 At 3° C warming, increased mortality burden of extreme heat will be in the tropics, where forecasts are currently worst



Note: Panel (a) shows the change in deaths per 100,000 population under 3°C warming in 2080-2099, inclusive of the projected costs and benefits of adaptation and income growth, using data from [4]. The tropics will bear most of this burden, and panel (b) shows that these are the areas where forecast quality—an essential adaptation—has been worst, reproduced from [3]. Decades of progress haven't closed the quality gap with higher latitude countries.

BOX 1.1

The History of Climate Services for Health

The first era was one of scientific evidence. From the early 1990s, health researchers began establishing how climate variability and change affect human health, spanning extreme heat, flooding, shifts in vector-borne disease, and food insecurity [19]. The argument was made [20] that global environmental change was reshaping population health at a scale existing frameworks could not capture, and this helped open a research frontier that continues today. Human health entered the IPCC's assessments beginning in 1990. Evidence accumulated, but it did not translate into operational decisions. Through this same period, sector-specific meteorology and public health developed separately: meteorological data were structured for forecasting and tailored (if at all) for agriculture, defense, and aviation, and rarely for health [21], while health systems organized their data around districts and service delivery. These divergent histories left deep incompatibilities that constrain integration to this day [10].

The second era delivered information but stopped at the dock. As the evidence base strengthened, demand grew for the meteorological sector to support health. The response took a form called the "loading dock model" [7], in which climate information was produced to meteorological standards and left for health users to interpret and apply, with little influence over its design. Information became technically available without becoming operationally useful [11].

The third era moved from delivery to co-production. The formal naming of climate services for health, catalyzed by the 2009 Third World Climate Conference and the 2012 Global Framework for Climate Services, brought climate and health actors together to design services around specific decisions [2]. Yet progress stalled at a familiar boundary. Most co-production kept climate information as the primary product with health systems positioned as end users, producing general risk interpretations rather than modeled health outcomes, and informing isolated decisions that were not embedded into health systems.

A fourth era is on the horizon. "Climate-Health Intelligence" is a phase focused on operationally integrating climate and health data into routine workflows and decision processes, enabled by rapid digitalization and AI across both sectors [22], and carrying forward the lessons of each era before it. The Weather Intelligence for Health approaches described here are part of this transition.

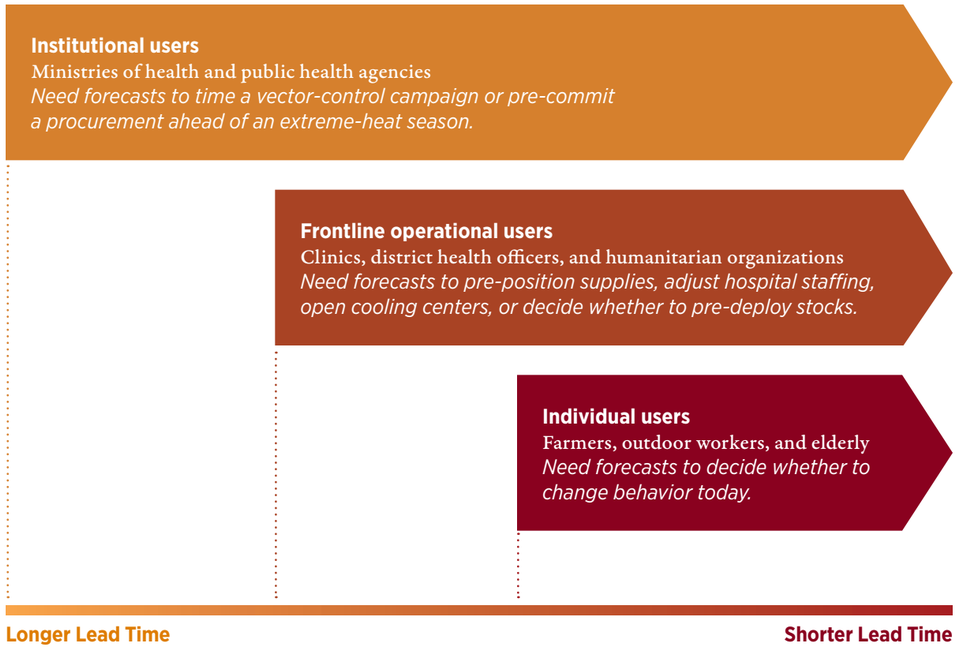
climate variability and change on health. They have grown from early climate monitoring toward systems that combine meteorological, environmental and epidemiological data and match it to specific risks such as heatwaves, vector-borne disease or poor air quality: seasonal forecasts feeding disease surveillance, early-warning triggers for preventive action, advisories tailored to a health decision. These services have evolved since the 1990s through four eras (Box 1.1), but a persistent set of constraints on forecasting capacity and access has meant the promise of CSfH has not reached as many people as could benefit. Availability of and access to forecasts were never the whole constraint; the institutional partnerships that carry information to a decision matter as much.

From supply to demand

Climate services have historically been built around what forecast producers can generate rather than around the decisions health systems need to make. The standard sequence starts with generating scientifically appropriate information, translating it into messages, disseminating those messages, and then working on how to get people to use them, and it is reinforced by a supply side that treats forecast accuracy as the main organizing goal [6]. That sequence puts the whole burden on the last step. Producing information and leaving it for users to collect has been called the "loading dock" model: technically and often freely available, without the salience, credibility or legitimacy that would make it usable [7]. The problem has been recognized, and difficult to overturn, for decades [8], [9], [10], [11].

A better approach should start instead with the decisions that need supporting, build the forecast and the system around them, and ask what evidence shows those decisions lead to benefit for people [12]. We use "weather intelligence for health" to mean forecasts, and the evaluation and institutions built around them, specified from the health decisions they serve. The WHO/WMO Climate and Health Joint Office is advancing a broader shift it calls climate-health intelligence, focused on operationally integrating climate and health data and information into support for health decision-making [13]. Building from the demand side up requires a set of specific people making specific decisions at specific lead times, most of whom the supply-driven approach has never consulted. Institutional users, such as ministries of health and public health agencies, need forecasts to time a vector-control campaign or pre-commit a procurement ahead of an extreme-heat season. Frontline operational users, such as clinics, district health officers and humanitarian organizations, need them to pre-position supplies, adjust hospital staffing, open cooling centers, or decide whether to pre-deploy stocks. Individual users, such as farmers, outdoor workers and older adults, need them to decide whether to change behavior today. All three levels matter for health, and each needs information at a different lead time.

Figure 1.2 Three levels of users



All three levels of decision-making matter for human health, and all three will require weather and climate information at different lead-times to inform different decisions.

In principle, the climate services practice of co-production should already have closed this gap [9]. In practice it has been hard to solve with numerical weather prediction forecasts—the traditional physics-based approach that revolutionized forecasting in the mid-20th century but that remains computationally expensive and technically complicated, and remains fully accessible only in the most advanced and well-resourced countries. A forecast producer who cannot change the forecast in response to what a user says can only consult those users and not build a forecast for them. The consequences of the supply-side approach are clear. When surveyed, 81% of national meteorological and hydrological services (NMHS) reported providing climate services for health, while only 23% of health ministries ran a surveillance system that used meteorological information [2].

Why now: weather in a changing climate

Better forecasts are needed now. Heat, heavy rain, storms, floods, droughts and the conditions driving vector- and water-borne disease already impose a large health burden concentrated in the tropics and falling on people who receive little actionable warning of any of it [1]. If climate change stopped today, the case for

building weather forecasts around health decision needs would remain strong.

But the climate is changing, and this also changes the value of forecasts. Health systems and households have always made anticipatory decisions from memory: the season that usually starts in June, the weeks that are usually hottest, the years when dengue usually surges. Memory of that kind is itself a forecast, and it works while the climate is stationary. Variability is rising, the events causing the largest health consequences increasingly have little local precedent, and even the thirty-year weather average used in seasonal outlooks and warning thresholds is being outrun by climate trends.

Adaptation will reduce the impacts, but it is incomplete. For heat, the best-quantified case, adaptive response cuts the projected mortality burden, with residual deaths concentrated in the low- and middle-income countries least able to afford adaptation [4], [14]. Those are the same places where forecasting is weakest. A forecast is anticipatory adaptation: it changes beliefs about the weather and so changes the decisions taken before it arrives [15]. It is also ordinary health protection, worth doing under any climate. Treating it only as adaptation ignores the present benefit and suggests the problem is a future one and not one of today.

"A lot of climate services failures are framed as these last mile service delivery challenges…but sometimes we do not take enough accountability in the first mile to say if we only built something useful, maybe someone would be more likely to use it."

Katie Kowal,
Data Science Institute, University of Chicago

Why now: AI is driving a second revolution in weather forecasting

The supply-driven approach was not chosen because it was best. At least three barriers made it the only feasible one, and with the advent of an AI-driven revolution in weather forecasting, each barrier is now collapsing [16]. For now, the skill improvement is at short to medium range, out to roughly fourteen days. Subseasonal and seasonal predictions remain at the research frontier, and we make the case for investing there on health decision value instead of demonstrated atmospheric skill.

The first barrier is cost. Until AI weather forecasts, running forecast models required supercomputers and infrastructure that excluded most national meteorological services in low- and middle-income countries (LMIC). The best physics-based forecasts run on machines costing on the order of $100 million or more (Box 3.1). Training an AI weather model from scratch still takes substantial computation, but the cost of producing an operational forecast from a trained model has fallen dramatically: in one case the energy cost of a single ten-day forecast dropped by roughly a factor of 1,000 [17]. More importantly, running, fine-tuning, downscaling and iterating on an existing model is no longer prohibitive for a national service. A forecast that once required a machine costing tens of millions of dollars can now, for many purposes, run on a laptop in minutes.

The second barrier was iteration time. Testing, adjusting, or altering a physics-based model made rapid user-focused development infeasible for all but the best-resourced services. If a health user's need implied a different forecast target, there was no practical way to build it, evaluate it and improve it on any useful timescale. AI weather models change this directly. By removing the specialist computational expertise needed to iterate, they could let anyone with meteorological knowledge build forecast output on a demand-driven cycle.

The third barrier was local tailoring: assimilating local data and adapting outputs to specific contexts, hazards, and decisions was prohibitive at scale. For years the only feasible approach was to build one forecast, push it downstream, and let users adapt. But the tailoring that sits on top of a forecast—bias correction, downscaling, and translation into a decision-relevant product—is now within reach of a modestly resourced national or regional LMIC service.

AI weather models relax these three constraints at once, and the value of local data shifts with them. Many national meteorological services in low-income countries hold the mandate to collect, steward, and share observations but have historically found few uses for their own records; some were sent to global repositories for better-resourced centers to run their models. Local data becomes far more valuable at the point where the service holding it can put it to work in a way it has previously been unable to.

This is what makes a decision-first, demand-driven approach newly possible. And it is why the status quo approach to climate services is now a choice rather than a necessity. However, the field of AI weather researchers has inherited the supply-driven framing for progress from their first-revolution forebears [16], so this choice is already heading toward a supply-side conclusion unless the paradigm shifts.

From last mile to first mile

Lifting the three barriers does not by itself change how climate services are designed. The challenge of getting information used is usually described as a last-mile problem: an excellent forecast is produced, handed to health agencies, and the forecaster concludes the job is done and the failure lies elsewhere. The last mile is real [11]. Dissemination, comprehension and action all break in ways a better forecast cannot repair, and closing that gap carries its own hard-won lessons. But the underappreciated break is at the first mile: supply-driven forecasts are not specified around user decisions in the first place, so what arrives is often not directly usable. Identifying those decisions, and benchmarking forecasts by their ability to support them, is what reduces the last-mile burden, because a forecast built around a decision is far easier to act on [2], [10].

Many barriers will be removed by the new capacities of AI weather forecasting, but a more capable technology delivered the same way will produce the same non-use. The way to mitigate this risk is to start from the decisions users need to make, which means the health sector states what it has to decide and when, before any forecast is built. Co-production has to

BOX 1.2

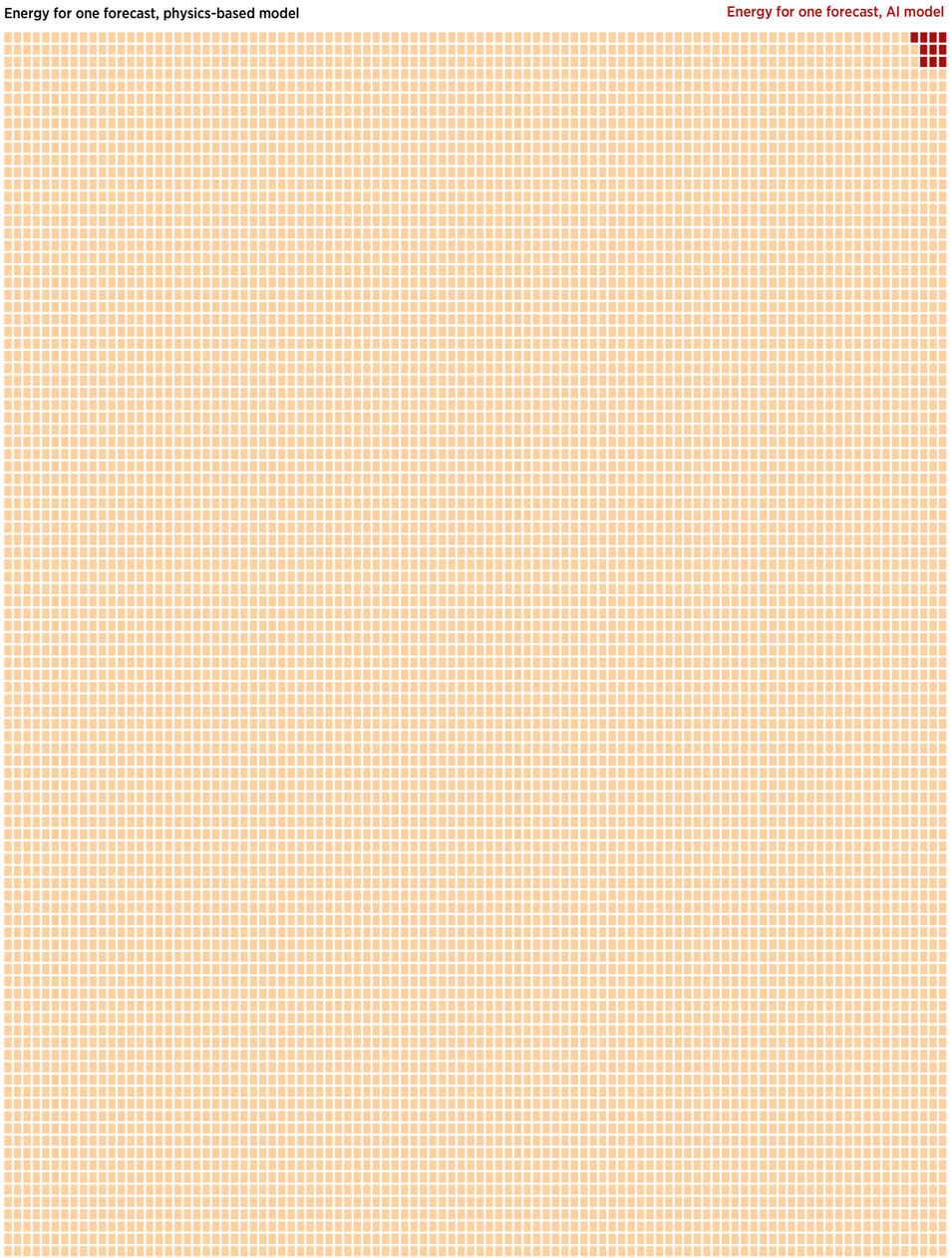
Two Revolutions in Weather Forecasting

Numerical weather prediction forecasts the state of the atmosphere forward on the order of days to weeks using the physical equations that describe atmospheric motion, energy transfer, and thermodynamics. The Navier-Stokes equations sit at its heart, drawing on a lineage of atmospheric science, thermodynamics, and statistical mechanics developed over centuries. Solving them recursively on large CPU servers has produced high skill in predicting atmospheric states. The first physics-based numerical forecast, run in 1950 by a team led by John von Neumann and Jule Charney, was among the earliest applications of scientific computing. Forecasts became operational in many countries, including the United States, by the mid-1950s [23]. That revolution advanced forecast skill mainly in the high-latitude temperate countries. It largely left behind the tropics, where the description of weather patterns and the data to capture them both lagged substantially behind the wealthier higher latitudes [3].

AI weather prediction began in earnest in 2022 with a set of scientific breakthroughs [24], followed within a year by models that surpassed the world's most accurate physics-based forecast on accuracy. The first operational AI weather forecast came online in 2025 [17]. These models do not predict from specified physical principles. They take data on the state of the atmosphere and, through transformer neural network architectures, iteratively predict subsequent states, learning from the physical behavior of the atmosphere itself instead of the rules of physics. The breakthrough rested on a consistent gridded record of weather at high frequency globally produced by ECMWF—itself a combination of observations and numerical prediction [25].

Training such a model requires heavy computation on GPUs. Once trained, producing a forecast costs on the order of thousands of times less computation than the NWP equivalent at comparable accuracy. AI weather prediction may be recent but it is already mature enough for operational use in meteorological services.

Figure 1.3 The same energy that produced one ten-day global forecast now produces about a thousand



Note: Each block represents the energy used to produce a single ten-day global forecast, comparing ECMWF's physics-based system with ECMWF's own AI model. Both are current operational systems. The figure counts the energy of producing a forecast, not the cost of the computer that produces it, and not the one-off energy of training the AI model. The trained model can be run by anyone with a single GPU. On elapsed time rather than energy the gap is smaller, about 26 times, because the physics-based system reaches 65 minutes by running hundreds of processors at once while the AI model takes two and a half minutes on one. More details in Box 3.1. Sources: [72], [73], [74].

happen at the design of the forecast, not as a delivery afterthought. A great deal of weather and climate information exists; what is missing is the specification that would make it answer a health question, and the structures that would carry it to the person making the decision at the appropriate time.

Starting from the decision instead of the forecast changes the paradigm of climate services. When a service begins on the user's side—identifying the decision that needs to be supported, then co-designing the forecast around it—it requires co-production of information, new institutional partnerships, clear mandates around issues like data stewardship and accountability for forecast errors, and benchmarking and evaluation of whether forecasts are accurately providing useful information for decisions and whether that information is benefiting the people it is meant to serve [11].

Who will benefit from the AI weather revolution?

Left to its own incentives, AI weather keeps adding skill where data are already dense, for uses it already serves well. AI weather models are trained and evaluated on data, and the incentives for the innovators in this field are to improve fastest where the data are densest in the observation-rich higher latitudes, not the data-sparse tropics where the health need is greatest. This might reinforce the forecast access gap that the first revolution in weather forecasting brought about, leaving behind billions of people in the lower-income countries in the tropics. What will determine the path that the AI-driven second revolution in weather forecasting will take is not just technical aspects of the models, but the decisions about how they are built, what data they use, how they are evaluated (especially benchmarking against existing reference forecasts), who they are designed to serve, and how they are governed.

The hard problems here, meaning data, benchmarking, coordination and governance, are not AI weather problems specifically. That does not make AI's contribution marginal. An accurate forecast is not the same as a good one, and neither is the same as a valuable one: value is set by the user's decision, not by fidelity to the atmosphere [18]. A producer can say nobody used their quality forecast, but this likely means an accurate forecast was made and not a valuable one.

AI's value is therefore conditional. It removes the constraints that made supply-driven climate services the only option, and could enable an operational decision-first approach for the first time. That value is realized only where the system around it is built: where data to train and validate models exist, and where governance structures deliver it to end users. Where weather risk is highest and the capacity to produce decision-first forecasts is thinnest, the potential value is largest and the surrounding investment most necessary. Chapter 3 sets out what the forecast was concealing, and the rest of the report is about the system that has to be built around it.

"The past used to be a usable guide. But variability is increasing, and what fills the information gap is forecasts"

Amir Jina
Harris School of Public Policy, University of Chicago

CHAPTER 2

Putting Health at the Heart of Weather Intelligence

AT A GLANCE

- Start from what health systems need to decide and when, not from what can be forecast now. A decision is a usable target only if five things can be stated: what the decision is, who owns it, what authority they hold, what resources they control, and what would change if the forecast improved. We call this demand articulation, and it is the first capacity a working system needs and the one most often missing.
- Setting nine health outcomes against six lead times gives fifty-four decision points, and four of them are operational today. The two outcomes those tour cells cover—heat illness and and injury from floods and storms—were already served by conventional disaster response. Where forecasting has reached health, it has reached the outcomes that look most like weather.
- Three levels of users—institutional, frontline operational, and individual—need different information at different lead times, and the delivery channel is part of the product, not just a detail of delivery.
- Two situations need different responses: forecast-to-action pipelines that exist and function poorly, and pipelines that were never imagined, often because the hazard is not recognized as a health risk so no one has asked for a forecast. The never-imagined ones concentrate where forecasting is weakest and where AI's benefit could be largest.
- The widest gap between what is recognized in national adaptation plans and what is acted on is at the subseasonal range: malnutrition is recognized in 78% of plans and only 20% describe concrete actions, heat in 75% and 22%, mental health in 44% and 5%. These are the never-imagined pipelines, not broken ones.
- Health outcome data are the hardest constraint. Where the climate-sensitive burden is heaviest, the record is weakest, and a field can only benchmark against what it can measure.

What health systems actually need

The question to start from is what health systems need to decide, and when, instead of starting with what can be forecast now. What the decision is, who makes it, and by when are what determine which forecast can support it. Putting the decision first also makes it possible to ask whether a forecast has value as well as quality: value is set by the decision, quality by how well the forecast reproduces the atmosphere [18].

Stating the decision is the first of five steps, and it is where the health sector begins and drives the process. The other steps are: who owns the decision, what authority that owner holds to act on it, what resources they control, and what would change if the forecast improved. We refer to this in the report as "demand articulation," and it is an essential part of making weather forecasts human-centered rather than atmosphere-centered. It is the first capacity a working system needs and the one most often missing, and where it is absent, nothing downstream can be specified: a forecast built for a decision that cannot proceed along all five steps is a forecast built for a label (i.e., "a health forecast") and not a forecast trying to inform a health target. It cannot be assessed for value, or benchmarked, because there is no stated action against which value could be measured.

Because forecasts are now easier to produce, we no longer have to build services around whatever happens to be forecastable. Driven by demand, we can start from what actually causes the health impact. Heat is the clearest case: a temperature-only forecast misses most of what determines physiological heat stress, since humidity, wind and radiation all matter, and the composite indices meant to fix this carry their own problems. WBGT requires inputs many services do not produce operationally, the Heat Index assumes shade and still air, and the Universal Thermal Climate Index needs radiation data that are rarely available at the point of decision [26]. The Jaipur case in Chapter 5 shows what that mismatch costs in practice.

Understanding health decision contexts

Climate information enters a health decision along a spectrum of integration [27]. At one end is a standalone climate-health bulletin presenting health-relevant conditions. In the middle are environmental

> "A met service and a health agency sit together; the met service asks 'What do you need?'; the health agency answers 'I don't know, what do you have?' The way through is not to start with the users but to start with the decisions."
>
> **Juli Trtanj, Trtanj Consulting**

indicators embedded in disease surveillance. At the far end is deep integration into routine health workflows, where climate information is built into the health system. Where a service sits on that spectrum determines how a change to the information would have to be applied. Integration context answers the third and fourth steps: where climate information already sits inside a health workflow, the owner has authority and resources and where it arrives as a standalone bulletin, the question of who can act on it is usually unanswered.

Early in the shift from supply-driven to demand-driven services, there may be only anecdotal evidence that information changes a decision, which can be supplemented with surveys, focus groups, or analysis of health outcome data. What matters is that evidence accumulates as the service runs. The chain runs from access, whether the information reaches the person at all, through engagement, recall and comprehension, trust and belief-updating, to behavior and then to health outcomes. Several of these links can be learned long before outcomes are measurable, and several are easier to detect than a shift in mortality or case counts.

Building the evidence that weather intelligence for health works

The chain above can be measured for effectiveness, but it rarely has been. Anticipatory action, early

> "It was like a loading dock. You build the forecast, put it somewhere on an FTP server, and the health folks will know what to do with it."
>
> **Hunter Jones,**
> **World Meteorological Organization**

warning systems, and climate services for health have generated a substantial literature on whether information reached people and whether they acted, and very little on whether their health improved [28]. That is not a criticism of the evaluations. Health outcomes are expensive endpoints, they need large samples and long horizons, and most programs could not have afforded them. It does mean the field is scaling on the strength of plausible mechanisms more often than measured effects.

Decision-driven services make this both more necessary and more tractable. More necessary, because the claim being made is specifically that starting from the decision produces better outcomes than starting from the forecast. That claim can only be settled by measurement, but the measurement can also lead to direct improvements in how messages are framed and communicated. More tractable, because a service specified around a stated decision has a stated action to measure against, which is what the supply-driven model never had.

Randomized evaluations and A/B testing can provide this evidence. Randomized trials compare districts, facilities, or communities that receive a service against those that do not, and staged rollouts make this possible without withholding anything from anyone who would otherwise have had it. A/B testing randomizes variants of the same product against each other: two thresholds, two channels, two lead times [29]. Where a full control group is not feasible, A/B testing usually still is. Several links in the chain are cheaper to measure than mortality and can be measured much sooner, including whether the information arrives, whether it changes what someone expects, and whether it changes what they do.

RECOMMENDATION R-E1

Evaluation of health benefits in the loop from the beginning, including assessment of communication, to aid the shift to decision-driven climate services

Most of this work will not be done by the people running the service, and most researchers who could do it are not connected to one. A standing facility that matches evaluation researchers to implementers, and that builds evaluation into programs already being funded rather than commissioning it separately afterwards, is a cheaper route to an evidence base than any individual study. It also changes what gets measured: three outcome classes matter, and they are usually collapsed into one. Whether people believed and understood the warning: whether they acted on it, and whether their health changed.

The cost of not doing this is not only that the evidence is missing. It is that without it, every subsequent investment decision is made on the same reasoning that produced the current situation, which is that the information seems useful and someone should therefore have it.

RECOMMENDATION R-E2

An evaluation and match-making facility, pairing researchers with implementers and building evaluation into existing programs

Communicating forecasts

Getting a forecast to people is not a single act. A heat advisory that reaches a district officer as a PDF attachment, a farmer as an interactive voice response call, a ministry as a dashboard feed, and a worker as a text message is four different products with four distinct kinds of potential successes and failures. Which one is being built has to be decided when the forecast is specified, not after it exists. Knowing which networks actually reach a given population, and whether anyone holds the relationship needed to use them, is as much a part of demand articulation as agreeing the threshold at which a warning is issued.

Where the formal channel does not reach, an intermediary sometimes does. Community health workers, agricultural extension agents, faith networks, and women's collectives already hold relationships that a new information service would otherwise have to build, and routing through them is often faster than building a channel from scratch. This pathway is much less developed than the volume of work on forecast production would suggest.

AI changes the economics of this step as well as of the forecast. Much of the work of integrating data, linking surveillance, and generating advisories can now be automated, including a layer that turns technical output into plain-language advice in a local language [30]. We do not go into detail on these capacities here. Many of them would deliver benefits even if forecasting technology stayed exactly as it is, and they are the subject of other work. What matters for this report is that improvements in forecast access are complemented by AI-driven improvements elsewhere in the pipeline: in data quality, in monitoring, in communication, and in what users can reach directly.

Health outcomes and scope

We consider forecasts that can address a defined set of health outcomes. Each has its own causal pathway from hazard to outcome and a characteristic biological lag, and that lag decides which lead times matter, and in some cases whether monitoring or forecasting is the decision-relevant input. For several outcomes, operational forecasting barely exists; for mental health, non-communicable disease and injury, there is almost no consideration of how forecasts could help at all. The causal pathways are described elsewhere and not re-reviewed here [1], [5], [31]. What matters for each outcome is the decision problem it defines.

Climate hazards and how they affect health

These outcomes are driven by hazards that forecasts actually predict: heat, heavy rainfall and flooding, drought, storms and cyclones, poor air quality, dust and fire smoke. A condition may have several drivers, and each hazard can be forecast over different timescales, with different accuracy, and by different techniques. That is why the decision, the driver, and the lead time have to be considered together, and it gives the structure for the matrix below. Three short examples show what a decision-first framing surfaces. None of the three is visible if the question starts from what a forecast can produce.

BOX 2.1

Monitoring v forecasting

Monitoring versus forecasting. Not every health decision is best served by a forward forecast [2]. For diseases that respond to environmental conditions weeks or months later, the biological lag itself supplies the lead time, and the valuable forecast might instead be a health-related forecast based on those inputs. If this is the case, monitoring of recent and current conditions can be more useful than a short-range forecast. But current-condition monitoring can itself be an input that improves short-range forecasting skill, and the strongest systems couple the two. We therefore consider both monitoring and forecasting as joint needs for public health, rather than framing decision needs on monitoring versus forecasting.

Vector-borne disease exemplifies this: the interval between favorable conditions and clinical cases spans the weeks needed for mosquito development and parasite incubation. A one-to-three-day weather forecast does little for a decision that plays out across a transmission season; monitoring of the conditions that create breeding sites, paired with a seasonal outlook, does far more. In Viet Nam, the D-MOSS system combines monitored earth-observation data, including a water-availability component that dengue models usually omit, with forecasts to estimate epidemic likelihood up to six months ahead, giving communities time to eliminate breeding sites before the season builds. In Niger, the monthly Climate and Health Bulletin analyzes current and expected climatic conditions to map areas of potentially high malaria transmission and route recommendations to control programs.

Niger's constraint is instructive: what limits the bulletin is not shorter-lead forecasting but gaps in vector and epidemiological data and in analytic capacity. For these diseases, the question is rarely whether the forecast reaches further ahead. It is whether current conditions are monitored well enough, and translated into action, at the lead time determined by biology.

A district hospital manager preparing for a hot season controls the roster, and rosters move on a two-to-four-week timeline rather than the two or three days at which heat alerts arrive. An outcome that looks solved from the supply side, because the forecaster produces a three-day warning, leaves the binding decision unserved: when the heat arrives, clinics may be understaffed relative to the risk.

A malaria program manager timing the year's spray round would gain nothing from a better five-day rainfall forecast, because the weeks of delay between rainfall and cases already supply the lead time they needed to act. What affects the campaign is a seasonal forecast combined with current conditions, which is how drug delivery is already timed across the Sahel.

A city health department deciding whether to send tomorrow's air advisory to registered COPD and asthma patients is making a small, cheap, repeatable call, telling people to pre-medicate and stay indoors. The forecast is available: daily global air-quality output reaches something like 200 million users, and alerts of this kind have been shown to cut hospital visits. But ground monitors are sparse and satellites have weak sensitivity to ozone at the level people breathe. The department can predict a pollutant and cannot verify whether anyone was exposed to one. Starting from the decision points investment at the observation.

Why timing matters

Weather and climate intelligence can support decisions across a range of lead times. Each is a different product with different current skill, and each connects to a different set of health decisions in the matrix (Figure 2.2)

- **Monitoring** of current and recent conditions, which is not a forecast at all, and which is the decision-relevant product for several health outcomes.
- **Nowcasting** of near-real-time conditions from 0 out to about twelve hours, using any method that draws on real-time observations. It supports acute, same-day responses such as heat-illness surge staffing or flash-flood evacuation.
- **Short-range forecasts**, from 1 to 3 days, support pre-positioning, staffing, and the opening of emergency shelters or cooling centers.
- **Medium-range forecasts** extend from 4 days to 10, or as much as 15, days and support procurement, campaign timing, and readiness decisions.
- **Subseasonal-to-seasonal (S2S)** prediction spans beyond 10 days or 2 weeks to 30 or up to 45 days for the subseasonal timescale, and into the overlap with seasonal forecasting of up to 2 months. This is the range where many health, food-security, and disaster decisions sit, and it has long been called a "predictability desert": historically the hardest range to forecast, and the least invested in.
- **Seasonal** prediction reaches from about 1 month to 6 months or up to a year ahead and supports resource allocation, stockpiling, and seasonal campaign planning.

Figure 2.1 Lead times for forecasts and the actions they might support.

What the matrix shows

The matrix sets the nine climate-influenced health outcomes against the six decision lead times, giving fifty-four decision points. Each cell describes a decision that could be supported at that combination and a status: operational today, available but limited, high value but not yet operational, low relevance, or monitoring being more important than forecasting. The statuses are not the results of a formal survey, but are judgments assembled from the existing literature (primarily [1], [2], [5], [26], [30]), interviews, and the June 2026 convening.

Five things stand out.

Four of fifty-four cells are fully operational. Four of the fifty-four cells are fully supported today, though improved forecasts could still provide benefits: short-range and medium-range heat warnings, and nowcast and short-range warnings for injury from floods, storms, and lightning. Both outcomes have almost no biological lag, so a weather forecast is the useful instrument, and both were already served by conventional disaster response. Where forecasting has reached health, it has reached the outcomes that look most like weather. Twenty-five cells describe capability that exists in prototype or in some settings but is limited in skill, integration, or coverage, which is a larger prize than the operational four and a different kind of work: scaling what exists rather than inventing what does not.

Capability clusters at both ends of the lead-time range and thins in the middle, which is an important timescale over which health systems could act. Almost nothing in the subseasonal column is operational for any of the nine outcomes, and few outcomes anywhere, especially in lower-income countries, have an operational subseasonal product. Four outcomes are marked high value and not yet available at that range, which is four of the six high-value cells in the whole matrix. Malnutrition is the exception, running on subseasonal timescales through food security systems that already work on that horizon. The monitoring and seasonal columns contain

> "The user changes quite a lot... sometimes its policies at the schools that need to close and the individual actually can't act there."
>
> **Kate Burrows,**
> **University of Chicago**

something for every outcome with no low-relevance judgment at all, while the nowcast, medium-range and subseasonal columns hold fourteen of the sixteen. Procurement, campaign timing, staffing, and pre-positioning happen in the middle of that range.

Mental health is already acted on inside packages built for other hazards, but not specified as a decision of its own. The three longest lead times are all marked high value and not yet operational, making mental health the only outcome with more than one such cell. WMO's 2023 survey records the lowest surveillance-integration rate of any outcome, at 3%, and 44% of national adaptation plans recognize it while only 5% act on it [31]. The gap is not that nobody has thought about it. Psychosocial support is already distributed inside packages built for heat, floods and cyclones, so it is acted on incidentally and never evaluated on its own terms. What is missing is a stated decision: no trigger, no threshold, no owner, and so nothing a forecast could be built against. That pattern is consistent with the wider deprioritization of mental health in health systems, and the forecasting gap follows the funding gap.

Vector-borne disease is where meteorological information is most deeply embedded, but also the outcome a better short-range forecast would change least. Vector-borne disease is the only row where monitoring is judged more decision-relevant than a forward forecast at two separate lead times, because the weeks between favorable conditions and clinical cases already supply the lead time a control program needs. Waterborne disease reaches the same conclusion by a different route at the subseasonal range, where the three-week anomaly windows used in cholera anticipatory action draw their warning from environmental and epidemiological lag rather than from a forward rainfall forecast. Vector-borne disease also has the highest surveillance-integration rate in the WMO

Figure 2.2 Health decisions and the forecasts that could support them

Each cell describes a decision that could be supported at that combination and its operational status. Read across a row to see how the decision changes with warning time. Read down a column to see which outcomes a given forecast can support.

Operational Today
In operational use in at least some settings, though often not where the burden is highest.

Available, but limited
Some capability exists, or a prototype does. Skill, integration or geographic coverage is limited.

High value, not yet operational
The health decision is specified. The forecast to support it is not operationally available.

Monitoring beats forecasting
A biological or environmental lag already supplies the warning time. A forward forecast adds little.

Low relevance
This lead time does not connect usefully to a decision for this outcome.

Health Outcome	Climate Driver	Monitoring	Nowcast	Short Range	Medium Range	Subseasonal	Seasonal
Heat-related illness	Temperature extremes; humidity (wet-bulb temperature governs physiological limits); urban heat island effects amplifying exposure	Heat-health surveillance	—	Activate heat warning	Stage readiness	Schedule clinical capacity	Pre-season resourcing
Respiratory illness	Air quality (fine particulates, ozone, pollen); dust aerosols; fire smoke; temperature; humidity (low humidity drives meningitis risk in the Sahel)	Air quality monitoring	Same-day advisories	Activate meningitis EWS	—	—	Vaccine stockpiling
Waterborne disease Cholera, diarrheal	Heavy rainfall and flooding (contamination of water supplies, sanitation breakdown); drought (water scarcity and quality decline); temperature; ENSO	Case and water quality surveillance	Boil-water advisories	Pre-position ORS, water kits	Alert facilities to caseload	Case-anomaly triggers	Pre-season stockpiling
Zoonoses Rift Valley fever, leptospirosis	Flooding and standing water; drought (increased human-wildlife contact at water sources); temperature (range expansion); land use change	Livestock and vector surveillance	—	—	—	Pre-position livestock vaccine	Time vaccination campaigns
Vector-borne disease Malaria, dengue, zika, chikungunya	Temperature (vector lifecycle and parasite incubation); rainfall and humidity (breeding habitat); ENSO (transmission season timing and intensity)	Surveillance supplies the lead	—	Forecast adds little	—	Anticipate transmission onset	Time vector control, chemoprevention
Malnutrition	Drought (crop failure, reduced food availability); flooding (crop contamination, destruction of food stocks); temperature (yield and livestock stress); ENSO	Food security surveillance	—	—	—	Pre-position therapeutic food	Pre-season nutrition response
Non-communicable disease	Heat extremes (cardiovascular events, acute kidney injury); cold extremes; ozone; pollen; temperature variability	Admissions surveillance	—	Alert hospitals, primary care	—	—	Season resourcing, procurement
Mental and psychosocial health	Extreme events causing displacement, trauma and bereavement; prolonged drought causing livelihood loss and chronic stress; heat; displacement and forced migration	Needs and displacement tracking	—	Deploy psychological first aid	Pre-position crisis teams	Plan for displacement	Pre-season MHPSS resourcing
Injury from accidents	Flooding and flash floods (drowning, trauma); tropical cyclones and severe storms; rainfall-triggered landslides; lightning; extreme winds	Hazard and injury surveillance	Flash-flood, storm warning	Order evacuation	Relocate patients, harden facilities	—	Season preparedness planning

Note: An online version that lists specific examples from the supporting literature and convening discussions, as well as the status of the evidence of health benefits, can be found at https://climate.uchicago.edu/ai-weather-forecasts-for-health/.

survey, at 14%. Attention has concentrated where forecasting adds least.

Where anticipatory systems do exist, almost none of them are triggered by health. In a review of 49 anticipatory action frameworks, 48 were built around a hazard and not a health outcome, one used a health trigger, and no early action protocol for health had been approved [28]. Where evaluations exist, they have generally measured food access, debt, and psychosocial stress. A lack of evidence on health benefits reflects the fact that health outcomes have rarely been the endpoint measured, not that the evaluations came back reporting no benefits.

RECOMMENDATION R-S4

Invest in monitoring and surveillance integration for climate-sensitive disease

Where the pipeline has not yet started

Two situations exist that need different responses. In one, a forecast-to-action pipeline exists and functions, well or poorly. In the other, the pipeline does not exist at all, because the hazard is not recognized as a health risk or because it is beyond current technical capacity. These never-imagined pipelines are not randomly distributed. They concentrate where forecasting is weakest and where AI's benefit could be largest.

The harder part of the never-started case is often not that the technology cannot produce the forecast, but that the hazard is not recognized as a health risk at all, so no one has asked for it. That gap is one of awareness rather than technology, and it is often the true first mile. It is especially acute at the subseasonal range: there is little forecasting capability there today, yet many decisions could be enabled on that timescale, and because people do not expect usable forecasts to exist there, they do not realize what decisions could be supported.

> "The data sharing system has been broken for a really long time. And we have worked around this for decades. Maybe we actually need to fix it."
>
> **Juli Trtanj**
> **Trtanj Consulting**

RECOMMENDATION R-D6

A subseasonal-to-seasonal (S2S) forecasting-for-health consortium

The subseasonal timescale is not uniformly free of articulated demand. Food security already runs on it: subseasonal indicators feed crop-failure outlooks, and therapeutic feeding supplies are pre-positioned on that horizon through systems such as FEWS NET. Where demand was identified around decision needs, the value of the timescale became clear. The demand exists in other health domains without the forecasts to support it. Two to four weeks of warning would let a health system stage its response to a heat episode and not merely react to it: schedule clinical capacity, mobilize and train staff, ready cooling infrastructure, and get advisories out before the week they are needed. The same window would let a malaria program anticipate a transmission season and move larval source management and spraying earlier instead of pinning them to a fixed calendar date. This is the same forecast horizon with a different history of demand. In a cross-sector scoping exercise in Nigeria, the preparedness actions participants named almost all required weeks of notice, while the products they asked for came in days or in seasons. Nobody asked for anything in the subseasonal range, and not because nothing there would be useful. Nothing in that range has ever been available to ask for (see Case 1: Nigeria).

The awareness gap is clear in National Adaptation Plans (NAPs) [31]. The outcomes with the worst gap between what is recognized and what is acted on are precisely the ones where medium-range and subseasonal forecast action could open new decisions: malnutrition is recognized in 78% of plans and actioned in 20%; heat in 75% and 22%; respiratory illness in 66% and 17%; and mental health in 44% and 5%. These outcomes are unaddressed but not unaddressable. Their pipelines have not yet been imagined. And the demand exists, most clearly in the plans health ministries write for themselves: across Health National Adaptation Plans (HNAPs), integrated risk monitoring and early warning is the second most-prioritized component of the WHO Operational Framework, at 81%, behind only leadership and governance. What is missing is the pipeline to deliver it, and a clear sense of what is being warned against.

RECOMMENDATION R-E3

Fund demand articulation for never-imagined pipelines

A pipeline can also stall on the health side. For several outcomes, and vector-borne disease most of all, the models that would turn a weather forecast into a case forecast have fragmented capacity across various research groups and are unvalidated against each other. WMO can reach NMHS, but there is no equivalent consolidated way to reach the climate-health modeling community to assess model accuracy and value, or to match that value to NMHS capacity. That is the gap the model-intercomparison recommendation addresses (Box 2.2).

RECOMMENDATION R-D4

Build a climate-health model intercomparison and improvement project

Health data limitations

Health outcome data are one of the hardest constraints for climate health intelligence. Where the climate-sensitive burden is heaviest, the record is weakest. That presents a challenge for bringing benefits of AI weather improvements to the populations that need them most. A field can only benchmark against what it can measure, and it will measure where measurement is possible.

In many cases, outcome data are not recorded at all. Roughly four in ten deaths worldwide are never registered, and across 133 countries registration ranges from 98% in the European Region to 10% in the African Region [32]. A widely cited global figure estimates roughly five million deaths a year associated with non-optimal temperatures, extrapolated from 43 countries with almost no data from Africa and none from rural areas [34]. Even where data exist at national level, that is rarely the resolution a decision requires, or the resolution AI weather can now supply. Mapping child growth failure across Africa at 5x5 km found countries on track nationally to meet WHO nutrition targets containing districts that will miss them badly [35].

Coding and case definitions vary between countries and within them, and heat deaths provide one extreme case of the challenges this creates. Most heat deaths are recorded as a cardiac event, renal failure, or as the consequences of a pre-existing condition. A construction worker collapses in 40°C heat, and it does not get reported as such. Even where death registration is complete, therefore, the count depends on the method. A model scored against death certificates and the same model scored against excess mortality are being asked different questions, and can rank differently on the answers.

Privacy and legal barriers differ qualitatively from the sovereignty concerns attached to weather data. Weather observations raise a question about who owns a national asset; health records raise a question about an individual and every privacy protection that implies. The practical effect is that data often cannot move between two arms of the same government, and only 14% of WMO Members have a formalized data-sharing agreement between the meteorological service and the ministry of health [2], a constraint confirmed by most participants in this report. Federated approaches address the legal barrier and are worth building, with their limits stated: they do not harmonize case definitions or eliminate re-identification risk [36], and they do nothing for records never collected. Identifiability is a property of combinations rather than of records, which is why weather-sensitive outcomes are usually the ones that survive protection. Governance is taken up in Chapter 6.

On the AI weather side, training and evaluating weather models against how well they reproduce the atmosphere was made possible by a globally standardized reference dataset [25] and a shared global benchmark allowing comparison across models [37]. Health has no equivalent. The Global Burden of Disease study is the closest and is a modeled synthesis, not an observational record, relying, where vital registration is absent, on verbal autopsy, a structured interview with family members used to infer a probable cause of death, and on covariate modeling. Evaluating an epidemiological model, a weather-intelligence-for-health model, or the effect of forecast

> "Why historically has cooperation between the meteorological services and the health sector not worked? Getting that right will help inform how you can take this model and apply it."
>
> **Titi Akinsanmi, Demistef.ai**

dissemination against those data risks scoring models against other models. A community-led effort to design a gold-standard reference dataset would begin to solve this. Resolution compounds it: estimates on a coarse grid can look accurate across a region while missing badly in the districts inside it, and an average that holds at scale can conceal populations getting no benefit at all.

RECOMMENDATION R-D1

A health-and-impacts gold-standard dataset, the analog to ERA5 for health impacts

None of these data challenges implies that the situation is hopeless or that we should not try to make progress on climate and health. Exposure-response relationships can now be estimated from weekly, monthly [38], and annual [4] mortality rather than daily records, which brings a large set of countries inside the evidence base with no new collection. The risk is not that the field waits for better health data to take advantage of the AI weather revolution. It is that it does not wait, builds benchmarks from whatever is at hand, and carries the geography of those data biases forward into the models that get selected.

Governance: the constraint is jurisdictional, not technical

Shared infrastructure does not on its own resolve the sharing problem. DHIS2 is deployed across roughly eighty countries, and the mandate to share across institutions is already written into how these systems are meant to work; what has not happened is the actual sharing. Success can make it harder, because data too fragmented to be valuable raises no question about who may profit from them, and assembled data raises it immediately (see South Africa case study).

Federated approaches, derived products, opt-in linkage, and aggregation to the spatial unit at which weather data is available are the near-term path. The limits of these fixes are clear: they do not harmonize case definitions, do not eliminate re-identification risk, and do nothing for records never collected. Identifiability is a property of combinations rather than of records, which is why the weather-sensitive outcomes are usually the ones that survive protection. The instruments to build on already exist (WIS 2.0, FAIR, the International Health Regulations, DHIS2) and updating them is faster and more durable than writing new frameworks alongside them.

BOX 2.2

Health Model Intercomparison and Improvement Project

Two people can need the same cholera forecast for entirely different reasons. A national program deciding whether to request vaccines from a global stockpile needs months of warning and a seasonal risk category. A district team deciding when to pre-position rehydration supplies needs weeks and a case threshold. Those are different forecasting problems, with different lead times and different tests of whether a model is good enough. Nothing currently tells either of them which of the dozens of available climate-sensitive disease models is fit for their decision, or whether a forecast improves on what routine surveillance already shows.

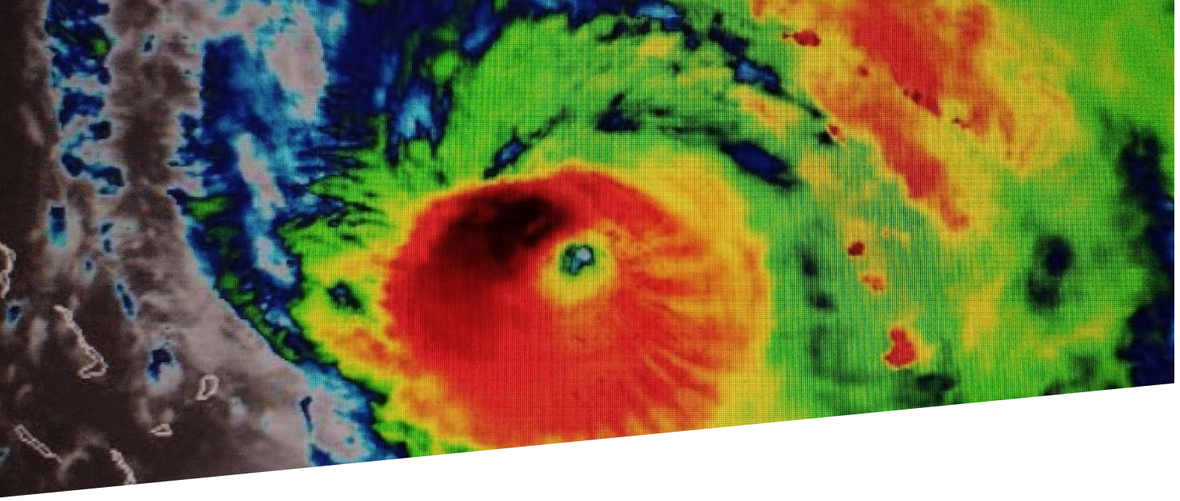
Other climate-sensitive fields that have faced this problem have built institutional infrastructure to address it. Since 2010, the Agricultural Model Intercomparison and Improvement Project (AgMIP) has run crop-modeling communities on shared protocols and standardized climate inputs under a neutral secretariat [39]. Across up to twenty-seven wheat models tested on the same experiments, disagreement between the models proved a larger source of uncertainty in yield projections than disagreement between climate scenarios [40]. Model choice was hiding uncertainty that no user could see. Health needs both halves of that comparison: which modeling choices drive the disagreement, and whether the disagreement changes a decision.

The field of climate and health is not starting from nothing. The Vaccine Impact Modeling Consortium (VIMC) coordinates disease-specific modeling groups on standardized inputs under a funded secretariat [41], collaborative forecasting hubs grew out of the influenza and COVID-19 challenges [42], and single diseases have been compared directly [43], [44]. None anchors comparison to a health decision. Even mature hubs do not provide triggers for action, and where comparability existed, it was constructed deliberately rather than found.

A climate-health model intercomparison and improvement project would supply that anchor: forecast targets, lead times, and evaluation metrics agreed with the people who would use them before any benchmarking is designed, supported by a coordinating secretariat, shared reference datasets where appropriate, and challenges that recur. Malaria and cholera are the places to start, chosen not for being best-modeled but for having decisions already waiting. Chemoprevention and spray campaigns are timed to the transmission season; vaccine campaigns and supply pre-positioning run ahead of outbreaks. Each has a lead time that matters, financing attached, and a program with the authority to act.

Two questions have to be settled with health institutions before any of this is designed. Which body defines the policy-relevant questions and creates demand for comparable outputs, as WHO and Gavi did for VIMC, given that modeling gets used where those relationships already exist [45]? And how to handle health reference data, the harder problem by far, where differences in case definitions, testing, and surveillance coverage narrow the possible evaluation metrics available?

Gains in AI weather skill are measured in the atmosphere. The climate-health model intercomparison project is the instrument that would measure how much value they'd bring to health, and let the epidemiological modeling community drive which forecasting gains are worth pursuing.



CHAPTER 3

What Weather Intelligence for Health Can Offer

AT A GLANCE

- AI weather models show accuracy improvements for many phenomena at the heart of climate and health and early warning systems—e.g., temperature, precipitation, flooding and tropical cyclones.
- Even the lowest-resourced NMHS can run a gold-standard global forecast on a laptop. Training a model is the most computationally intensive piece of the AI weather production chain, but once trained, the models can be run cheaply.
- Pre-trained global models can be improved with techniques on a continuum of resource requirements—light-touch post-processing, bias correction, blending, and full local training. Several of these are within reach of a modestly resourced national or regional service today, and costs will decrease.
- Downscaling may be the most consequential application, because what it replaces is among the most expensive things a weather service can run. The binding constraint has moved from computing power to obtaining the fine-resolution training targets, which often do not exist where need is greatest.
- The cost collapse changes which data are worth collecting. Observations that were never worth the price of assimilation into a numerical model are now cheap to try, which makes the value question urgent.
- Extremes are the gap that matters most for health, because the events that trigger health decisions sit in the tail of the distribution or outside the record entirely, and current models appear to trade missed events for fewer false alarms.
- Where weather or health data are weak, the models are weak. AI does not create co-production where none exists, does not answer whether a warning is understood, and does not repair institutional relationships.

New capabilities for weather forecasting

AI weather models have moved from early experiments less than five years ago to operational systems that on some metrics surpass the best physics-based models, at a fraction of the energy and computational cost [46]. That capability arrives into a large gap in forecast production capacity. In 2021, only 60% of national meteorological services in the Southern African Development Community region ran or even tested a numerical weather model in-house, and about one in five produced downscaled seasonal or climate-change projections [47]. Capital investment in observing and computing infrastructure has often outpaced the operational budget to sustain it, leaving hardware without the capacity to exploit it [48].

Producing accurate global forecasts domestically is now achievable for many low-income countries, because the models are pre-trained and a producer does not need the computational power or expertise to train one from scratch. Downscaling those outputs to local resolution is increasingly affordable, and where local data exist, bias-correcting against local observations is straightforward. Iterating quickly, whether by blending models or bias-correcting, lets a producer extract far more value from the same underlying models. Fine-tuning on health-relevant variables and benchmarking against them is within reach, as is automating the translation of forecasts into plain-language advisories. Several of these are already running operationally in well-resourced services and remain out of reach elsewhere. Training a model from scratch is possible but still resource-intensive, and likely beyond most national services at present.

These capacities change what local data is worth operationally. The mandate to produce, accredit or validate a forecast, held nationally or regionally, is what turns local data and local understanding into the highest-value forecasts. Forecasts could be produced nationally, regionally or privately, but much of the value added at national level comes from the mandate for data stewardship on the weather side. Owning that data, or being able to access it, is what enables the opportunities above, and it has many uses at once: ground truth, benchmarking of historical forecasts, initializing forecasts, bias correction, and downscaling. Even a record with gaps is worth more as forecasting turns data-driven, and its uses are widening.

BOX 3.1

The costs of weather forecasting have collapsed

A single forecast from the world's most accurate physics-based system consumes an estimated 14.4 kWh, produced on a dedicated supercomputer procured for roughly €80 million and re-procured every four to five years [72]. Training AIFS, ECMWF's own AI model, took about a week on 64 GPUs. Producing a ten-day global forecast from that trained model takes roughly two and a half minutes on a single GPU [17].

On computation time and energy, the two models differ by more than an order of magnitude. On wall-clock time, the AI model is about 26 times faster: roughly 65 minutes against 2.5 minutes. On energy, it uses on the order of a thousand times less [73]. The gap between the two ratios is the interesting part. The physics-based system reaches 65 minutes by running hundreds of processors simultaneously, all drawing power for the full hour; the AI model reaches 2.5 minutes on one GPU. Speed measures how long a forecaster waits. Energy measures machines multiplied by time, and on that measure the collapse is two effects multiplied together.

The asymmetry matters more than either ratio. Matching the physics-based system requires every serious adopter to build or buy a supercomputer. The AI capability is trained once, by one well-resourced institution, and reused by anyone with a single GPU. For most of the populations this report is concerned with, the realistic alternative was never the same forecast produced the conventional way. It was not an independently produced forecast at all. Sovereignty has always been an issue in this field; AI is what makes it achievable, because a country that can run and adapt its own models owns the process rather than receiving its outputs. Because model training is the most resource-intensive component, but is performed infrequently, pooling these resources across services would provide considerable benefits.

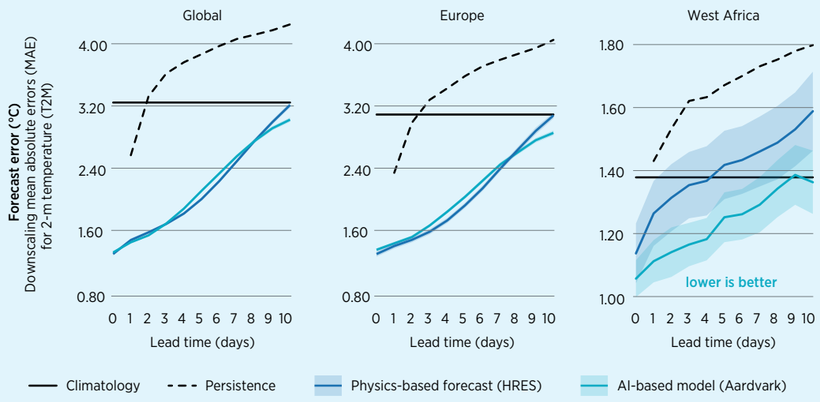
What that opens is not one cheaper forecast but a different operating reality entirely. A service that could previously afford one model and one run can now run several, more often, with more ensemble members, for less than it used to spend on a single deterministic pass. The binding constraint moves off compute and onto the things this report argues for: benchmarking, translation into decisions, and delivery to the people who need them.

BOX 3.2

Training AI models directly from observations

Aardvark Weather [64] is the first system to replace the full numerical weather prediction pipeline with a single machine learning model. It takes raw satellite radiances, station reports, ship observations and radiosonde profiles, estimates the state of the atmosphere without data assimilation, and produces both global gridded and local station forecasts, with nothing from an operational center required at run time. It ingests roughly 8% of the observations conventional systems use, matches or exceeds the Global Forecast System on most variables, and generates a full forecast in about a second on four GPUs, against the order of a thousand node hours that assimilation and forecasting require at ECMWF. Over West Africa and the Pacific, its local temperature forecasts beat a bias-corrected IFS baseline at every lead time out to ten days, and those are the regions where that baseline is closest to what agencies actually run.

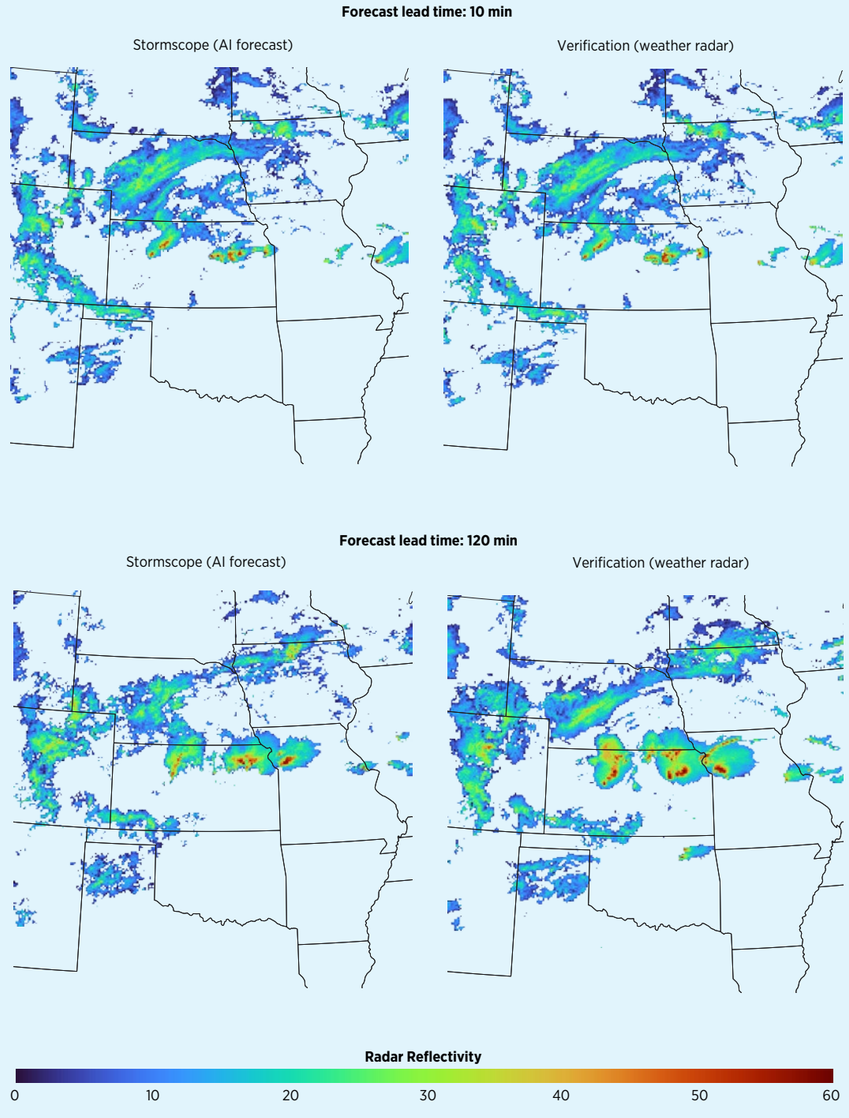
Figure 3.1 Models trained from observations show greater accuracy over tropics



Note: The Aardvark model [64], trained directly from observations, matches the best physics-based forecast (HRES) for accuracy globally up to 10 days of lead time for temperatures at the surface. Lower error values are better. Forecasts below the horizontal climatology line mean that they outperform using daily average for that time of year. Strikingly, the areas where current physics forecasts perform worst, like West Africa, show the greatest improvements with observation-trained AI models, giving a useful temperature forecast out to 10 days.

Stormscope [63] applies the same logic at storm scale, where the observational problem is hardest. Conventional convection-permitting forecasting requires assimilating dense radar networks into kilometer-scale physics models, infrastructure that exists in a handful of countries. Stormscope learns storm evolution directly from sequences of geostationary satellite imagery, with radar as an option to improve accuracy, and produces forecasts at 6 km and 10-minute resolution with no assimilation at any point. Trained over the continental United States, it outperforms the operational High-Resolution Rapid Refresh model on precipitation skill for the first three hours and stays competitive to five, using single-member forecasts rather than the ensemble averaging earlier systems needed. A twelve-hour forecast runs in about seven minutes on one GPU. A satellite-only configuration trained in the tropics has not yet been tested, and if the skill transfers, the low local-data requirement could open extreme precipitation forecasting in highly vulnerable countries. Both systems are evaluated entirely on atmospheric accuracy. Neither tells us whether the forecasts support a health decision, yet.

Figure 3.2 AI models can match the computationally-intensive physics-based standard model for one of the most difficult forecasting tasks: near-term extreme rainfall



Note: The Stormscope model [63] is trained only on radar and publicly-available geostationary satellites. The figure shows that the AI model can closely match radar observations out to 2 hours. At lead times of up to 2 hours, it performs as well as or better than the best operational physics based model over the United States (the High Resolution Rapid Refresh model) at a tiny fraction of the computational cost, or produce a probabilistic ensemble of tens of thousands of forecasts instead of a few dozen. Given the high human and economic toll of extreme precipitation, lowering the barrier to good forecasts could provide substantial benefits.

"AI now simply lets you do things that weren't even conceivable of before."

Niall Robinson
NVIDIA

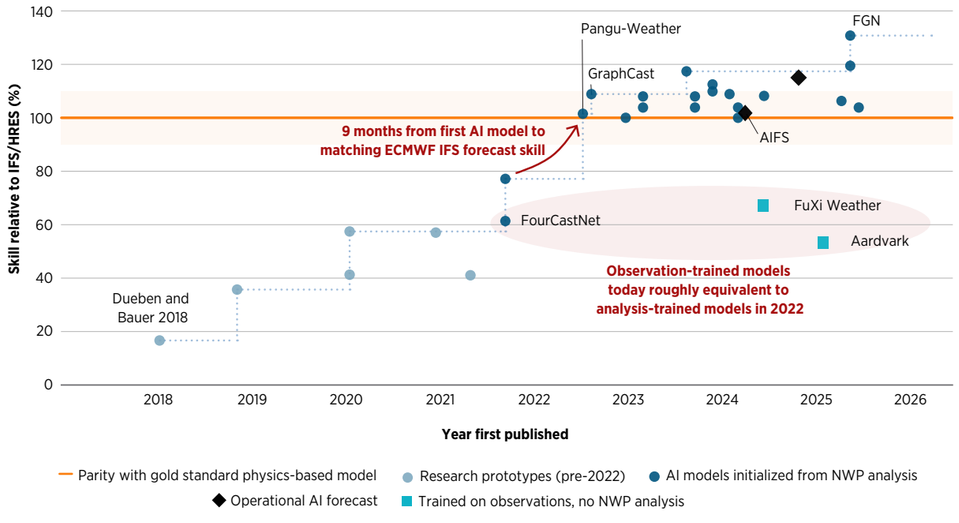
Beyond temperature and precipitation, a number of AI models for specific hazards have become operational. AI cyclone models now predict formation, track, and intensity competitively with the best conventional models, and one has entered the US National Hurricane Center's operational workflow. Deep generative nowcasting outperforms conventional extrapolation on precipitation from zero to three hours [49], [50]. AI flood forecasting achieves at five days' lead time what a leading global system achieved at zero, and now runs free and without registration across more than 150 countries [51]. These hazards drive several health outcomes at once: flooding drives waterborne disease, injury, malnutrition, and the psychosocial consequences of displacement.

The models are still a step behind where they would need to be for greatest health relevance. Cyclone track is largely solved and rapid intensification is not, and translating a track into surge, inland flooding and wind at a human-relevant scale is further behind still. Nowcasting skill is learned from dense radar networks that exist in a handful of countries, so the satellite-primary route is the version that could reach the tropics, and it has not been tested there (Box 3.2). For floods the riverine problem is close to solved, while urban flash flooding, storm surge and inundation depth remain gaps. One evaluation is already measuring whether any of this changes outcomes: a randomized trial in Bihar found that an accurate, free, publicly available flood forecast reached almost nobody until a human dissemination layer was built and tested [52].

The AI weather technological frontier

This section presents some possible directions for AI weather developments as expressed by leaders in this field. While none of these is explicitly tied to or affected by

Figure 3.3 The first revolution in weather forecasting brought one extra day of useful forecast lead time per decade. The second revolution has matched that skill in four years



Note: Early prototypes reached 20% of the accuracy of the operational physics-based forecast. The first AI models arrived in early 2022, and by late 2022 the leading models had passed the operational physics-forecast. ECMWF's own AI forecast became operational in February 2025. Models that matter most for forecasting sovereignty are the newest and the furthest behind: trained on observations alone, they need no analysis from a global center, and they are roughly where the analysis-initialized models were in early 2022. Metric shown is root mean square error of 500hPa geopotential at 3-day lead time. Each point may vary by about ±10%. Data are from WeatherBench 2 [37], updated 2026.

demand from health services, they are illustrative of the type of general progress, and the rate of that progress, that the field expects to continue.

Within one to two years, the frontier is forecasting directly from observations. Producing an operational forecast from raw observations, at skill comparable to the best physics-based system, removes the data-assimilation step that has been the most expensive and most centralized part of the conventional pipeline. AI data assimilation and early subseasonal capability sit in the same window. For a national service that has never had assimilation infrastructure, this is the change that matters most, because it removes a dependency rather than adding a capability.

RECOMMENDATION R-D7

Fund training from observations, including satellites, for data-sparse and tropical regions

Within five years, two things arrive together. The first is on-demand forecasting, combining analysis data where they exists with the latest available observations, so that a forecast is generated when a decision needs one instead of being produced on a fixed operational cycle. The second is on-demand downscaling in both space and time, where resolution becomes a request made of the model. One expert put the five-year target at forecasting across all timescales at sub-kilometer resolution.

Within ten years, it is possible we could see a unified Earth system foundation model capable of forecasting across weather, climate, air quality, hydrology, and extreme events at multiple spatial and temporal scales. The more ambitious version of the same horizon is direct forecasting of weather-driven impacts, not weather variables themselves, at resolutions down to tens of meters, which is the technical form of the health-first model discussed below.

Forecasting in the tropics and other data-limited regions has been a challenge for decades. Several AI weather experts suggested that over the next ten years, the developments most likely to transform capability there are extreme-event prediction, downscaling and probabilistic forecasting, with lead time and computational efficiency ranked lowest. A health decision-maker choosing between a two-day and a three-week warning would rank differently, and the gap in skill at the subseasonal range, where procurement, campaign timing and pre-positioning decisions are made, suggests a different priority. The distance between those two rankings is the argument for health-driven incentives to push forecasting into the subseasonal range.

A few developments would greatly aid low-income countries, and each is a case where funding could materially shape the direction of progress, because all are highly useful and unlikely to be rewarded by frontier-pushing publication. One is a movement toward open-source models, tools and frameworks, particularly from the private sector. This is a necessary condition for gains in data-limited regions, because open-source models let operational users benchmark, tailor, combine and generate forecasts for their own needs. Another is novel data sources that predict weather impacts rather than weather variables, which change the relevance and value of forecasts. And transfer learning, meaning robust foundation models that learn from data-rich regions and fine-tune to data-poor ones with minimal local data, could deliver large skill gains in the most data-constrained countries.

Despite these potential developments, bottlenecks remain: advanced computation and the expertise to use it; trust in AI systems; and, most concretely, that adapting and fine-tuning pre-trained models remains ad hoc with mixed results, so robust methods for fine-tuning with reliable gains are themselves a research gap. Software platforms for adapting, retraining, and validating models are emerging but still demand too much time, effort, and expertise. A reliable agentic AI that a non-expert could interact with to execute those tasks is another opportunity that would enable weather intelligence to benefit health—a "de-experting" operational tool.

"The magic of these things is actually learning from the global picture, and then you're calibrating and post-processing with the local knowledge."

Katie Kowal,
Data Science Institute, University of Chicago

A range of improvements requiring different resources

Not every producer can train models on local or global data, but many improvements need no training at all. They sit on a continuum of resource requirements. Light-touch post-processing recalibrates an existing output without touching the underlying model. Bias correction adjusts a forecast to match local observations. Blending runs several models and combines them. Full local training builds or retrains a model on local data. Downscaling to higher resolution can sit at several points along that range. Each works at a different stage of forecast production.

Blending shows what the lower-cost end now makes possible. Instead of training from scratch or reading an off-the-shelf output, a service can run several existing models itself, calibrate them against a decision-relevant baseline on local data, and blend them to outperform any single one [53]. This is newly feasible because inference is cheap: a national service can run four or five global models on operational timescales, which it could never do when it was at best running one.

Choosing between them requires knowing what each is worth, which means forecast quality has to be transparent. That need grows with AI weather models, and it implies someone must be mandated to steward quality. Benchmarking, meaning transparent evaluation of models for a specific purpose, tells a producer what accuracy a model has for a given use and therefore which improvement is worth its cost. Chapter 4 takes this up.

One gap matters more for health than for most applications. AI models learn the statistics of what they were trained on, and the events that trigger health decisions sit in the tail of the weather distribution or outside the observational record entirely. A model trained without Category 3-5 cyclones could not forecast a Category 5 storm, failing to extrapolate from weaker events to stronger [54], and NWP outperforms AIWP on some extremes [55]. Current models appear to issue fewer false alarms at the cost of missing events, which is the wrong direction for warning decisions where the cost of acting is low relative to the cost of failing to act [56].

Two approaches address this. Weighting the training objective toward the tail makes a model attend disproportionately to rare events, at some cost to average-case accuracy. Transfer learning exploits translocation: where extrapolation fails, a model can learn an extreme where it is common and carry that skill to a region with too few local occurrences to learn from [54], [57]. Which of these delivers benefits for health decisions, for which hazards, and at what cost is unlikely to be settled by the frontier research agenda because improving operations does not merit a major academic publication in the way that improving global accuracy would.

RECOMMENDATION R-D5

Build shared tools for adapting and evaluating forecast-improvement techniques across the resource spectrum, benchmarked against health-relevant extremes

What is possible now, and what may become possible

Some capabilities are operational today; others are plausible within three to ten years. Health-first models fine-tuned to output a health-relevant target directly, more skillful subseasonal prediction, direct training from observations, and unified models combining air quality, climate and weather may all become not only possible but affordable. That trajectory reshapes the resource picture: an approach that is costly today, such as training a model from data, can become routine and inexpensive, migrating from the heavy end of the range toward the light end. This is especially so if demand from the health sector spurs development around capacities that matter for health decisions. Planning should account both for where an approach sits now and for where it is likely to sit as the field matures, holding the resource continuum and the time horizon as distinct axes.

Downscaling, personal, and hyperlocal forecasting

Downscaling is potentially one of the most important applications of AI weather, because what it replaces is among the most expensive things a weather service can run. A kilometer-scale regional model needs its own compute, its own boundary conditions and its own data assimilation, and that assimilation step alone is highly costly [58]. A generative downscaling model bypasses it. CorrDiff, trained to produce 2 km fields over Taiwan from 25km reanalysis, learned that mapping from three years of data and runs on a single GPU at a small fraction of the energy of the regional model that produced its own training targets. The same divergence appears at global scale: roughly three orders of magnitude in energy against a factor of tens in computation time. The model synthesizes radar reflectivity, a field absent from its inputs entirely, which for a service without radar is an entirely new capability, not a cost-saving. It is well past proof of concept, having beaten an operational 3 km model in China for accuracy out to three days [59] and being implemented by ECMWF inside the open Anemoi framework [60].

Dynamical downscaling has been effectively out of reach for African institutions, and that barrier may already have fallen for anyone who can obtain the training data. Obtaining them is now the binding constraint. A downscaling model learns a mapping from coarse input to fine truth, and the fine truth has to come from an operational regional model, a regional reanalysis, or a dense observing network. All of these exist where forecasting is already good, which makes conventional downscaling the clearest instance of the inheritance we are arguing against: the cheapest capability is now available and needs the one asset the places that need it most do not have.

Several approaches may overcome this in the medium term. Model-agnostic downscaling is furthest along: a single diffusion downscaler, trained once and applied to different AI and physics-based models, has shown improvements at short-range lead times [61]. Learned Earth embeddings used as downscaling predictors, in place of hand-encoded terrain and land-use covariates, may provide skill where there are no stations to learn from. Off-the-grid downscaling would predict at arbitrary points rather than mapping one gridded product onto another, which matters wherever a health decision sits between grid cells. A practical caution for health is that how far a forecast can be downscaled is bounded by what it is downscaled from, and the input may introduce further error [62]. In a health system, a false alarm is how a warning stops being believed, which is the rationale for health-specific downscaling.

The sovereignty argument is usually made about owning data and running a global model. But owning a downscaling setup is closer to owning a local physics-based model at a fraction of the cost, and may be enough on its own. The global model, and the training of it, need not be locally owned. The model-agnostic approach makes this concrete: the global forecast becomes a swappable input, while the tools that make it locally meaningful stay with the country.

> "What we want to really see in Africa is to have more Africans on board with this in this technology. Not to be a user, but really to be on the driving seat."
>
> **Ousmane Ndiaye, ACMAD**

A further frontier points toward the individual. As downscaling sharpens and inference gets cheap, forecasting and warning can move to the person: individualized impact forecasts coupling environmental predictions with physiological models to produce a person-level risk score, for example a one-to-five heat-risk score built from simple prompts about activity, clothing and location. Sensors and wearables may soon allow tailoring to individual risk and activity. The forecasting capacity for this may exist soon, and several concerns come with it. Consumer wearables measure activity well and are poor at core body temperature and dehydration. They are also likely to reinforce unequal access, because universal adoption is unlikely, which could itself inhibit scaling. Keeping dissemination public and accessible is a mandate question taken up in Chapter 6.

Data and forecasting sovereignty

AI is a route to greater forecasting sovereignty: a shift from sending data to global centers toward running forecasts on local data. Observation-primary models can already produce skillful forecasts from satellite data or ground observations alone [63], [64], lowering dependence on global assimilation centers. Models trained elsewhere do not necessarily transfer to the tropics without retraining, and local data is what lets them transfer and gain skill, which is itself an argument for holding that training and observation data. Regional bodies may be a better home for retraining than any single national service, because training is costly but infrequent and differs from operational running, and because many weather needs span borders, so coordinated effort across a large domain yields more to each country than working alone.

RECOMMENDATION R-D10

Fund regional model-training and localization centers

New realities of data in the AI weather era

As forecasting turns data-driven, what a producer can do with data changes, and so does which data are worth collecting. Low-cost and non-traditional sources have long been proposed: sensor networks [65], rainfall inferred from cell-tower signal interference [66], pressure and temperature readings from mobile phones [67]. The barrier was economic, not physical. Under conventional numerical weather prediction, assimilating a single observation can cost the equivalent of running 30 to 100 full model forecasts, and operational systems assimilate only a small fraction of the observations available even from well-calibrated sources [68], [69]. Against that fixed cost, a noisy uncalibrated reading from a phone was not viable, and the documented biases of such sensors widened the gap. The question was never whether the data were good, but whether they were good enough to justify the cost of getting them into the machinery at all. For most opportunistic sources the answer was no, which is why crowdsourced data has tended to enter through post-processing or verification and not at the assimilation stage [70].

That calculation is likely to have now flipped. Data-driven models built to ingest sparse, irregular, off-the-grid observations directly, replacing assimilation with a learned encoder, remove the expensive optimization loop that a new observation used to be weighed against [64] at roughly three orders of magnitude less compute and using a fraction of the input data, and data do not need to be always perfect or always online. These properties matter most exactly where WMO-grade networks are sparsest and non-traditional data most abundant: much of the tropics and much of sub-Saharan Africa.

The cost collapse changes the incentives rather than the outcomes. It becomes cheap to try adding a source and cheap to be wrong about it, but the actual value of cell-tower rainfall, crowdsourced pressure, or a low-cost sensor network for a particular forecast, region and lead time remains almost entirely unmeasured. Under numerical weather prediction, the high cost gave the answer: a source not obviously worth dozens of forecast-equivalents was left out. This is a research frontier with real promise, which makes the value question more urgent than it has been until now.

In this case, benchmarking becomes an important operative tool. If any cheap source that localizes is worth having, the way to find out which ones actually localize, and by how much, and for which decisions, is to test them against the forecast improvements they produce. Decision-relevant benchmarking is what turns a newly affordable experiment into knowledge: it distinguishes the non-traditional sources that measurably sharpen a heat warning or a flood nowcast from those that add noise.

RECOMMENDATION R-D8

Fund research into non-traditional data sources for localization and downscaling, benchmarked against health-relevant metrics

Cost was not the only constraint. Meteorological services will generally not assimilate observations that fall outside established standards without a WMO directive or supplementary guidance, so even a source whose value was demonstrated could not be used. The fix here is cheaper than an instrument. An opportunistic observation becomes operationally usable when the type of sensor it came from is known and recorded, which makes this a metadata requirement rather than a hardware one, and one that fits inside existing infrastructure like WIS 2.0 and FAIR.

Three uses of weather data

For decades, the question asked of any new weather observation was whether it could improve a numerical weather model. Assimilating a low-quality source was expensive and the gain marginal, so most were refused. That test no longer governs, and the answer now depends on which of three questions is being asked. Data serves three purposes, and each requires different properties.

Treating data as having a unitary purpose is why cheap data gets rejected wholesale and expensive data gets over-specified. It also gives the pull toward convenience over need (Ch. 4) a mechanical explanation: evaluation carries the strictest requirement, so it drifts hardest toward wherever high-grade data already exist. That bias is structural, which makes it harder to argue away and

Figure 3.4 Three uses of weather data in the AI era

easier to design against. Every data recommendation in this report can be run through one test: what is the use, and does the data have the properties that use requires? Two consequences follow immediately. First, novel and low-cost data do not solve the evaluation problem, so the cheap recommendation cannot substitute for the expensive one. Second, a dataset optimized for benchmarking may be a poor training set—and vice versa.

Combining with other models

Many other models matter for health: human physiology, and models of vectors, pests and disease transmission. AI weather models can be coupled with these, and can feed features of their predictions into a forecast model's loss function. One appealing shortcut is to fine-tune a weather foundation model to output a health target directly, e.g., dengue cases and not just standard weather variables, skipping the intermediate steps. It is a promising direction, but it is not yet known whether a health-first model degrades the weather model's meteorological skill and coherence.

The simpler near-term path is to couple AI weather models with a downstream physiology, vector or transmission model, which may yield more skill. This is a question of benchmarking against health outcomes, and the returns are largely to operational health needs and not to publishable research, which is why it needs dedicated support.

RECOMMENDATION R-D3

Decision-oriented benchmarking of forecasts for health

Scenario planning for compound and rare events

A different frontier can be transformative for modeling compound and rare events. Compound-risk modeling is fundamental to decision makers and currently impossible to do soundly, because to date it would have involved extremely large probabilistic ensembles of forecasts to find the probabilities of rare events, such as when heat

and flood coincide. Cheap inference makes very large ensembles affordable, so distributions once sketched from a few dozen members can be populated by thousands.

A more novel approach inverts the question. Rather than running the atmosphere forward and waiting for a dangerous combination to appear, a weather service, health agency or research group can specify the outcome and ask what produces it: how likely it is, what conditions precede it, and how far ahead those conditions announce themselves. Generative models that sample from a learned climate, in contrast to the typical approach of stepping forward in time make this tractable, and can be run backward from a specified state to generate what came before it [71].

This answers a different question than a forecast does. A city writing a plan for a hazard combination it has not yet experienced does not need to know when the event will arrive. It needs to know what the event looks like, how often to expect it, and whether anything gives warning. Where a precursor proves identifiable and far enough ahead, the plan gains a trigger, and the exercise yields a monitoring target. The capability is still developing. With the correct incentives and applications, this capacity may develop quickly into a planning aid as climate risks shift.

RECOMMENDATION R-D9

Fund generative scenario models conditioned on health-relevant compound events, benchmarked on precursor lead time

From scenarios to decisions

Weather scenarios are only half of what a decision-maker needs. The other half is what available responses cost under each of them. The probability distribution from an ensemble forecast is a decision-ready object, because a decision made under uncertainty needs the probability of crossing a threshold, not a best guess about what will happen. The economics of decisions under uncertainty are well established. Acting early costs something and is sometimes wasted; not acting risks a loss that is usually much larger. The probability at which acting becomes worthwhile is the ratio between the two [56]. For most health warnings, that ratio is low, because opening a cooling center or pre-positioning oral rehydration salts is cheap relative to the deaths avoided, which means health decisions are often triggered by probabilities that would look unconvincing in any other setting.

Ensembles were expensive, so they stayed small. Fifty members is a typical operational ensemble because each member is a separate run on a supercomputer. Cheap inference removes that limit, and a distribution that previously had only a few dozen members can now easily have thousands. Low-probability tails, which are where many health-relevant thresholds sit, become statistically estimable with few assumptions.

The second kind of scenario focuses on the response, not on the weather. Given the range of what the atmosphere may do, what happens if a district pre-positions supplies it may not need, against what happens if it waits and the outbreak arrives, against what happens if it acts at a lower threshold and does so four times a season? These are questions about the health system, not about the atmosphere, and answering them needs the forecast distribution joined to what the response costs and what it prevents. Large ensembles enable these existing methods to be fully utilized. Scenarios of this kind are also what allow a service to be designed before it is trusted: a health ministry can see what the system would have told it during past seasons, and what each response would have cost, before committing to act on it in a live one.

A district officer weighing a warning is weighing a career risk as well as an outbreak, and being handed a single number to act on offers no defense when the event does not arrive. Being able to show what each health response option cost across the full range of outcomes is a different position to be in, because the choice can be justified by the reasoning that is involved in choosing those actions.

Despite the promise of these cheap ensembles, a caveat on present capacity should be noted. Current AI weather models tend toward sharp, confident forecasts, which score well on conventional accuracy measures and are a liability at exactly the low-probability thresholds where health warnings are triggered [56]. A model that is reliable in the middle of the distribution and overconfident in the tail will systematically under-warn for precisely the decisions this report is concerned with. Scenario tools are only as good as the calibration underneath them, which makes this a benchmarking requirement.

BOX 3.3

AI weather's carbon footprint

AI weather models are not carbon-free, and this needs to be considered alongside any efficiency claim. The largest energy use occurs during training: published estimates for energy demand during the training stage across seven major models range from around 400 kWh for the leanest (FourCastNet) to roughly 25 MWh for the most demanding (NeuralGCM) [74]. But model training is a cost paid once, and the model is reused by anyone who can run a single GPU, which is the structural difference that matters more than the raw kWh gap. Physics-based models have no equivalent training stage (though they do have development costs), but every forecast they produce carries its own energy cost: ECMWF's own accounting puts the energy cost of producing a single 10-day forecast with the physics-based IFS at about 14.4 kWh [72]. Once trained, an AI model produces that same forecast for a fraction of a kWh. Across the seven models studied, even the least favorable case consumes 21 times less energy than IFS over a year of routine operation, and the best case over 1,000 times less, a few Wh. On carbon footprint specifically, even the highest-emitting AI model stays below IFS's footprint, when producing one 10-day forecast daily for one year, by at least a factor of 6.

Weather forecasting is not exempt from a pattern researchers have flagged across AI more broadly: efficiency gains at the level of one model do not automatically translate into lower total resource use once adoption, ensemble size, and model proliferation are counted [75]. We can see that dynamic inside weather AI already. Training a probabilistic, ensemble-capable version of an AI model costs several times more than its deterministic predecessor [74], and the field is moving toward exactly that kind of model. Commentators covering AI's footprint more generally have made the same point: reported efficiency gains are often real but incomplete, because they rarely account for how much more frequently the technology then gets used [76]. Broader skepticism about AI's climate promises generally is a reasonable prior to bring to any single-sector efficiency claim, including for weather [77].

The AI models are "greener" in a sense, but the counterfactual is unclear. Many of the NMHSs who might adopt them are currently not running any energy-intensive NWP, so the comparison is not AI against IFS but AI against nothing. That makes the more useful comparison for this report not "AI vs. physics-based, same output." It is what becomes possible once the marginal cost of an additional forecast, ensemble member, or localized regional product drops to a fraction of a kWh. A meteorological service that could previously afford one model and one run can now run several, more frequently, with more members, within the energy budget it used to spend on a single deterministic pass with NWP. For most of the populations this report is concerned with, the realistic counterfactual was never "the same forecast, produced the physics-based way." It was not an independently run forecast at all. The binding constraint shifts from compute and carbon to the parts of the system this report actually argues for: verification, translation into decisions, and delivery to the people who need them.

The carbon footprint of AI models is an important public policy question, but in the domain of weather forecasts it certainly needs to be considered against the public good that can be achieved by protecting lives and livelihoods and improving health. This is yet another reason why centering future forecast development around decisions that matter in the real world, not an ever escalating competition for accuracy alone, matters - so that higher energy costs for training new models aren't wasted.

What the forecast was hiding

For most of the populations this report concerns, the forecast has been a real barrier. Too coarse to say anything about a district, too late to act on, or built around a variable that does not track what makes people ill. Removing that barrier is a substantial achievement, and this chapter has been an account of how it is being removed.

When a climate service did not change anything, the explanation was usually found downstream: the message did not arrive, or arrived and was not understood, or arrived and could not be acted on. Some of that was true. But a diagnosis located at the last mile is a diagnosis that asks nothing of the people who produced the forecast, and it survived partly because it was comfortable. Often the forecast was the problem: specified around the wrong variable, at the wrong scale, at a lead time no decision could use, or simply absent for the decision at hand. As the forecast stops being the binding constraint, that explanation stops being available, and what it was holding in place becomes visible. Six things become visible, and none of them is new.

1. **The decisions were never specified**. Services were built around hazards, seasons and administrative calendars rather than around a stated decision with an owner, a threshold and an action attached. This is why the health decision matrix has fifty-four cells and four operational ones, and why the exercise of filling it in is harder than it looks: for most outcomes at most lead times, there is no stated decision to write down.
2. **The evaluation metrics were never matched to the health effects**. Heat is the clearest case. Physiological stress depends on humidity, wind and radiation as well as on air temperature, and the plans that say so most clearly still trigger on forecast maximum temperature, because that is what forecasts provide. The gap between the variable that predicts illness and the variable that arrives is a specification problem, and it was invisible while nobody could have supplied the better variable anyway.
3. **Lead times were never aligned to decisions**. Procurement, staffing, campaign timing and pre-positioning run on weeks. Forecasting invested in days and in seasons. The two-to-four week window that most health system decisions actually need is the least developed range in the field, and demand for it was never registered because nothing in it was ever available to request.
4. **Ground truth was never collected**. AI weather models could be trained and evaluated fast because everyone agreed on a single reference record. Health has no equivalent, and the closest thing to it rests on modeled synthesis exactly where vital registration is thinnest. The same weakness runs on the weather side: these models learn from reanalysis, reanalysis is only as constrained as the observations beneath it, and where station density falls away, the training target is driven more by model physics than by measurement. The Lancet Countdown, the field's principal monitoring effort, reports that scarcity of health and climate data across Africa leaves many of its global indicators with limited regional coverage. The most-cited multi-country analysis of temperature and mortality drew on around 74 million mortality records across thirteen countries [14]; a later one drew on around 400 million records from 41 countries [4]. Neither includes a single African country.
5. **Institutions never had to speak to each other**. A supply-driven service needs a producer and a recipient. A decision-driven service needs a health ministry able to state a decision, a meteorological service able to build for it, and an agreement about who holds the data, who issues the warning, and who is answerable when it is wrong. Where those arrangements are broken, they hamper both dissemination and the joint production cycle that a decision-first service depends on. Nothing forced anyone to fix them while the forecast was the reason the service was not working.
6. **Health benefits of forecasts were rarely evaluated**. Whether a warning is understood and acted on is a behavioral question that a more accurate forecast does not answer, and whether acting on it changed anyone's health is a question the field has rarely asked. This is not negligence. Health outcomes are expensive endpoints, and a service that could not specify what action it was meant to trigger had nothing to measure against.

Read as a list, this is discouraging. Read in sequence, it is not, because each of these was intractable for the same reason.

Specifying decisions is pointless if no forecast can serve them. Matching the metric to the health effect is pointless if the better metric cannot be produced at the scale and lead time the decision needs. Aligning lead times is pointless if the missing lead time is unreachable. Collecting ground truth is a large investment with no return if the model it would improve cannot be retrained for years. Building institutional arrangements produces nothing until there is something worth transmitting through them. Evaluating a service is not possible until the service has a stated action to evaluate. Solving one of these alone, without an improved forecast, would have produced no benefit. That is why they persisted, and why they persisted together.

The difficulty of solving them has not changed. Ground truth is as expensive to collect as it ever was, mandates are as hard to negotiate, and evaluation is as slow. What has changed is that the return on addressing them is no longer zero, because the thing that made every one of them futile has moved. This is also why the problems are most acute, and most worth attacking, for the populations the first forecasting revolution passed over. For them, every link in the chain is weak at once, which is what made incremental repair pointless and what makes coordinated investment worth more now than the sum of its parts.

The chapters that follow take these up in turn. None of them is a story about forecasting.

> "Anybody can establish an AI model today. We're going to be overwhelmed by models. But at the end of the day, it's not about the models. It's really about the health of a human being."
>
> **Ousmane Ndiaye, ACMAD**

Capacity realignment through climate-services training

Every capability in this chapter assumes someone can operate it. The scarce skill is not model development, which will stay concentrated in a small number of centers, but operational use: knowing which model to run for which decision, how to calibrate it locally, how to read what it says, and how to explain that to a health officer. That is a realignment of existing meteorological expertise, not a replacement of it, and staff who have been through this training describe feeling further enabled. Training forums, summer schools, and embedded fellowships inside meteorological services and health ministries are the instruments, and they are also the cheapest demand-generation mechanism available, because people who can run these models start asking for things they would not otherwise have known to want.

RECOMMENDATION R-U1

Capacity realignment through AI-weather training forums, summer schools, and embedded fellowships

CHAPTER 4

The Quality Mandate: Why Trusted Forecasts Depend on Benchmarking

AT A GLANCE

- The health record against which a forecast would be verified is thinnest exactly where the climate-sensitive burden is heaviest, which is why ground truth has to be fixed before anything else can be measured.
- As forecasts get cheap, anyone can produce one, and the old basis for trusting a forecast (that only a handful of well-resourced centers could make one) no longer holds.
- There is therefore a mandate for quality, and it is a different mandate from the one that existed before: a transparency standard, owned by some accountable body.
- Forecast value for a health decision is not the same as forecast quality in the atmospheric sense; benchmarking against the decision can rank models differently than atmospheric accuracy does.
- Benchmarking does not just measure models; it selects them: it determines which models get improved, adopted, and used. Benchmarks built on global accuracy steer model development toward the data-rich regions that dominate those metrics, while benchmarks built on local, decision-relevant skill steer it toward the places with the greatest unmet need.
- A benchmark is a public instrument, which is why open models, or at minimum an archive of hindcasts, matter. A forecast that cannot be independently rerun cannot be independently checked.
- Where reference data are scarce, benchmarking drifts toward what is available, not what is important. Closing the ground-truth gap is what makes decision-relevant benchmarking possible.

There is no standardized approach today to validating a forecast or disclosing its quality against decisions or decision use-cases. As the cost of producing forecasts falls, and more models become accessible and producers of forecasts proliferate, that gap becomes more consequential. A mandate has to exist to keep quality transparent and known, with an accountable body able to say, transparently, what a quality forecast consists of for the uses that matter. That is the quality mandate.

> "Evaluate against the actual decisions you're trying to improve, not the model."
>
> **Hunter Jones,**
> **World Meteorological Organization**

Three questions called benchmarking

Benchmarking compares forecasts against each other, against a reference forecast, and against what was later observed, using a standardized and repeatable test. Three different questions get asked under that name, and the argument that follows depends on keeping them apart.

Atmospheric accuracy. How closely does the model reproduce the state of the atmosphere, scored against the historical record and averaged globally? This is what the public AI weather leaderboards measure and what model developers optimize against.

Skill at the health-relevant threshold. Does the model correctly call the days when the variable a health decision turns on crosses the value that triggers action? A heat action plan that activates on wet-bulb globe temperature, because that is what maps to physiological stress, is asking a narrow question. On the days that mattered, in the districts that mattered, did the forecast get the crossing right? A model can be excellent on the first question and poor on this one, because global average error is dominated by ordinary days in data-rich places.

Skill at the health outcome. Does the model help predict dengue cases, heat admissions, or cholera presentations? This is the question a health service actually has. It is the one that decides whether adopting a model changed anything.

The second and third are both decision-oriented, and both differ from the first. The third is the one to build toward. The second is the one most countries can build now.

The distance between them is a data problem and nothing else. To score a model on dengue cases, a dengue case record is needed: of known quality, at the resolution the decision uses, over enough years to evaluate against. Where the climate-sensitive health burden is heaviest that record is thinnest, and in many places it does not exist at all. To score a model at a health-relevant threshold, two things are needed. The threshold, which the health sector can state today. And weather observations, which are imperfect but present. That asymmetry is why the near-term work looks the way it does.

The example in this chapter is the second of these. The benchmarking presented below evaluates models at health-relevant thresholds on meteorological variables. It does not evaluate skill at predicting health outcomes, because the outcome data to do that at scale does not yet exist. This is the tractable form, and it is not a weaker version of the same idea, because the threshold comes from the health side and the decision still sets the test. What it cannot tell us is whether a model that calls heat thresholds well also supports a heat mortality decision well. Closing that gap is the purpose of the health-and-impacts reference dataset we recommend (R-D1).

Building benchmarks around decision-relevant criteria, meaning the variables, scale, lead times, and baseline a decision actually requires, is not a simple change of metric while otherwise following the same practice. Those criteria are not a technical choice available to a forecast producer working alone. They come from demand articulation, and the health sector supplies them: which decision, at which threshold, with how much warning, and what it costs to be wrong. A benchmark set without that input is a meteorological benchmark with a health label on it. This runs in both directions. Producers have to understand the decisions their forecasts are meant to serve, and health users have to understand what a forecast can and cannot tell them, because a benchmark score that nobody on the health side can interpret cannot inform a decision either.

A benchmark is also a public instrument: something others can pick up and rerun against new models as they arrive, rather than a private evaluation issued once and taken on trust. That is what makes benchmarking a tool

BOX 4.1

Forecast quality is not forecast value

In 1993, the meteorologist Allan Murphy asked what makes a weather forecast "good," and found that the single word hides three different things.

The first is **consistency**: whether a forecast says what the forecaster actually believes. A forecaster who judges the chance of rain at 20 percent but issues 40 percent, to hedge against a verification score or to avoid calling yet another dry day, has broken that correspondence, whatever the weather does next.

The second is **quality**: how closely the forecasts match what is later observed. This is the subject of conventional verification; Murphy's point about it is that quality is never a single number. Accuracy, skill, reliability, resolution, sharpness, and discrimination each capture a different facet, and a forecast can be strong on one and weak on another.

The third is **value**: the benefit a forecast delivers to someone making a decision. Forecasts, in Murphy's account, possess no intrinsic worth; they acquire value only through the choices they change. Value depends on the decision, on what it costs to act and what is lost by failing to, and it differs from one user to the next even for an identical forecast.

Quality and value do not move together. Murphy showed that two forecasting systems can trade places when ranked by accuracy against decision value: the system with the lower average error is sometimes the less useful one for the decision at hand. An accuracy score used as a stand-in for usefulness can therefore select the wrong model. That is the conceptual root of the reversal shown in this chapter, and the reason a health decision has to be benchmarked against the decision itself and not against atmospheric accuracy alone.

for accountability, and it is why open access to forecast models, or at least to an archive of hindcasts, is essential for ensuring quality [78]. A forecast that cannot be independently rerun against the standard cannot be independently checked against it, and a model that cannot be checked in the decision-relevant setting where it would operate cannot be responsibly deployed there, however well the model claims to perform. The quality mandate therefore depends not only on the right criteria but on the openness that lets anyone hold a forecast to them.

The changing landscape and the new mandate

The old basis for trusting a forecast rested on three things: an understanding of the physics underlying the model, evaluation of those models against observational data, and on scarcity. The relative scarcity of models played a role: only a handful of national or international centers could produce a global forecast. In many cases of operational dissemination, particularly in LMICs, evaluation was not performed on local data due to data availability or quality issues. Some trust in the model came from the fact that it may be the only model available. Scarcity lent authority. That has changed. The cost of producing a forecast has fallen and just like with NWP, AIWP forecast quality can vary. So the question of how quality is maintained has to be answered again, from scratch.

The quality-and-transparency mandate obtains new importance because of the proliferation of models. The full set of mandates a working system needs is set out under governance (Chapter 6), which, because national situations vary so much, names the capacities that must exist without prescribing which agency holds each. Many of the others exist to build the enabling environment that lets AI forecasts actually deliver a benefit, while the quality mandate establishes whether the forecast has value for health decisions in the first place.

Benchmarking doesn't just measure models, it selects them

Forecast value for a specific purpose is not always aligned with forecast quality in the atmospheric sense [18], and decision-relevant benchmarking can produce substantially different model rankings from those conventional accuracy measures produce. To benchmark for health value, health practitioners and decision-makers should define the decision thresholds and the metrics; local observations should act as ground truth; and social and public-health scientists should work alongside meteorologists in the process.

An example of the difference between a meteorologically focused model evaluation and a decision-relevant benchmark for a heatwave may help with concreteness. Pinning a heatwave to within a degree at a specific hour can matter less than flagging that a danger threshold will be crossed, with two or three days' lead time to act. The right decision criterion depends on knowing what the decision actually is—who needs to make that decision and what actions will it cause—and it can tell a very different story about a forecast's value from a story about its quality. Better models do not automatically translate into better outcomes. This is true for weather models [56] and for health ones [12]. We work through an example of benchmarking health-relevant forecasts of extreme heat in Figure 4.1.

The operational question

Decision-relevant benchmarking shifts model selection efforts toward health and social science, which raises an operational question. If the health sector defines the metrics and the ground truth, who runs the models, validates them, and develops them over time? Health services do not do this today. Operational infrastructure across this pipeline is one of a set of capacities, each needing a mandated home. Where a country has meteorological capacity to realign, running and benchmarking the models may sit naturally with the NMHS, increasing the importance of its role, while the benchmarking criteria are set jointly with the health decision-makers. That placement is not universal, and it is likely to vary on a country-by-country basis.

Benchmarking as a mechanism for trust and sovereignty

Decision-relevant, transparent benchmarking is how a nationally or regionally mandated body can maintain quality control and public trust as the number of available models grows, and how it can adopt forecasts on its own terms. The current global benchmarks used to score AI models [37] may not reflect the local, decision-relevant skill that effective use requires, because they are built on global atmospheric accuracy. Benchmarks built on global accuracy steer model development toward the data-rich regions that dominate those metrics, while benchmarks built on local, decision-relevant skill steer it toward the places with the greatest unmet need. Local, decision-relevant metrics have long been discussed in the evaluation of NWP models in theory [18] and in practice [79]; the greater operational need now is to bring an alliance of meteorological and health researchers to bear on assessing the skill of AI models, so that assessment is not based on atmospheric accuracy alone. That steering is the point: it is how investment in the development of these models can be brought to serve users in low-income countries and ensure that society does not replicate the current status quo of either climate services or high-income-country forecast accuracy.

RECOMMENDATION R-D3

Decision-oriented benchmarking of forecasts for health, against both weather and health outcomes

Benchmarking will be most effective as a model-agnostic approach. Every model can then demonstrate what skill it has for a given purpose, which is why open-access, open-source models (or at least models that can be benchmarked so their skill is transparent) matter as a public good. The example of the health-relevant heat forecast benchmarking in Figure 4.1 shows that models may rank highly in different countries for the same purpose. Model rankings may also change across different health decisions. Recurrent evaluation and benchmarking, run on a shared decision-relevant framework and shared local data for ground truth, let competing models, including those inside platforms like DHIS2 (District Health Information Software 2, the open-source health-information platform used by many ministries of health), be verified against evidence rather than producer, vendor, or private company claims. It rests on a plain fact: there is no single globally best model, for AI or for conventional NWP. Different models do better on different things in different situations, and much of that is knowable only by benchmarking for the purpose at hand. The key change from current practice is to identify the problem first and then find the model that serves it, rather than starting from a model and looking for a problem to fit.

RECOMMENDATION R-D11

An open-source or benchmarkable standard for AI weather models

Figure 4.1 Benchmarking forecasts using a health decision-relevant threshold for extreme heat

A Choosing a decision-relevant definition: in this example, the heat-mortality relationship drives the selection of the local 95th-percentile temperature as a threshold and the three-day lead time

The event is defined by a local percentile, not a fixed temperature. Mortality is flat through the middle of the temperature distribution and steep in the upper tail. Three days is an actionable lead time for decision-makers. A conventional meteorological benchmark might look at average error in the forecasts. A health decision-maker cares more about the probability of detecting days in this dangerous range in order to issue a warning. Once such a day is detected, small errors in the temperature are less likely to matter. For this reason, we compare a model ranking based on the more meteorological approach of average error to one based on probability of detection. Heat mortality is just one example. Other health decisions would require different variables at different lead times.

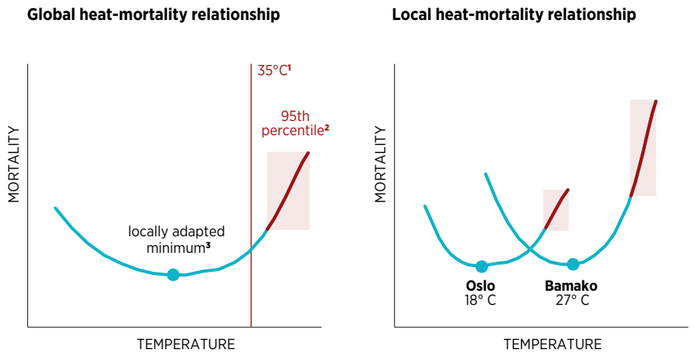
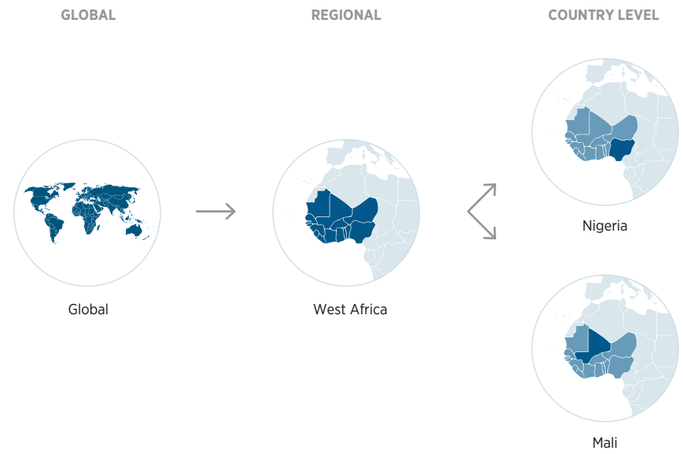


Figure Notes

1. A fixed 35°C threshold measures who lives somewhere hot.
2. A percentile measures who is outside what they are adapted to locally. Hot temperatures in this range have large effects on mortality; missing one of these days can be deadly.
3. The locally adapted minimum can be very different from one location to the next; it is near 18°C in Oslo and near 27°C in Bamako [4].
4. Forecast error at the locally adapted minimum comes with little consequence for health.

B Benchmarking at the scale relevant to the decision

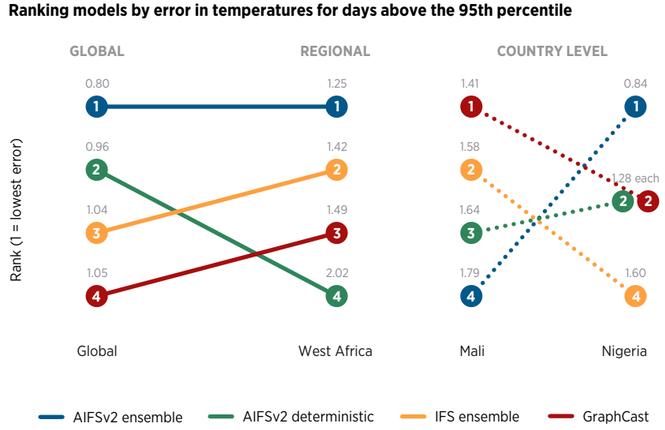
Global benchmarks are the current norm. Benchmarking closer to the scale at which a decision needs to be made is essential.



Mali and Nigeria are two countries inside the same benchmarking region, so a single regional score is all a shared benchmark could report for both of them.

C Ranking the models the conventional way: average forecast error over the same days

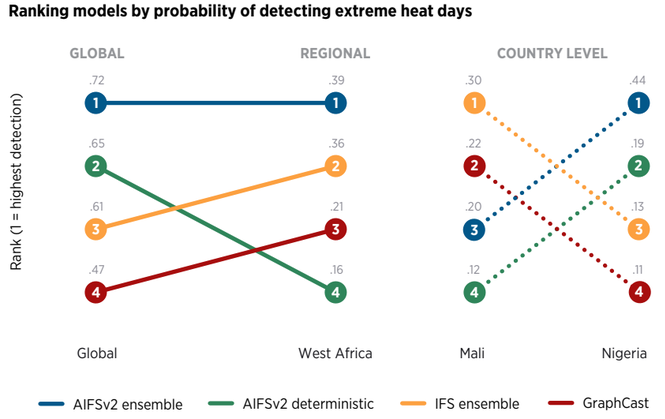
When assessing AI weather models, a measure of error is a conventional approach. In this case, lower error (root mean square error, RMSE) is better. This is the ranking a meteorological service would inherit from the existing AI benchmarks.



Globally, the four models are close together on average error. Within countries they spread out and differentiate themselves, with clear winners and losers, so optimizing for the global score does not seem to carry down to the local one. The question for a decision-maker in one place is whether the best model there is the same as the best model somewhere else. Averaging over a large area hides the disagreement. Where the decision is actually made, the two ways of ranking choose different models. In Mali the order is close to reversed. The model with the lowest error there is the one this metric ranks last globally. Average error cannot separate two models that it scores at 1.28 K each in Nigeria, even though one of them detects nearly twice as many dangerous days.

D The best model globally is not always the best model locally for a specific decision

For health decisions, the probability of detecting extreme heat days is in many cases a better decision-relevant metric than the average error in temperature on those days, because it is the detection of an event that triggers a warning. This would not be the conventional benchmark, but it is the health decision-relevant one. Driving model choice by looking at health decision-relevant metrics leads to different conclusions about what a "good" model is.



Ranked by detection, the order matches the error ranking globally and across West Africa but not within countries, where Mali and Nigeria pick different best models. Global and Nigeria have similar ordering. All four models match between those two columns, so a ministry following the leaderboard would have chosen well. Mali is where the same ranking is furthest wrong. The two models the world ranks first and second come third and fourth there, and Mali's best detector is the global third choice. West Africa matches neither country, so the regional score is not a compromise between its two parts.

BOX 4.2

Building the ground-truth record in the age of AI weather: ENACTS.

Enhancing National Climate Services (ENACTS), developed by Columbia University's International Research Institute for Climate and Society in partnership with national meteorological and hydrological services, is the most fully realized existing model for the observation-based ground-truth record we recommend building [81], [83], [84]. Since its 2012 launch in Ethiopia, it has been implemented in more than twenty countries, nationally and through the regional climate centers in East and West Africa, producing gridded rainfall and temperature records at 4-5 km resolution, spanning forty years for rainfall and over fifty for temperature.

The method resolves a tradeoff that otherwise forces a bad choice. Station observations are accurate but sparse, unevenly placed, and in many national networks declining in number; satellite and reanalysis products are spatially and temporally complete and freely available, but carry systematic, terrain- and season-dependent biases. ENACTS uses the station record, however thin, to compute climatological adjustment factors, applies those factors to the complete satellite or reanalysis field, and then merges in observed station values wherever they exist. The output inherits the coverage of the proxy and the accuracy of the observations.

What makes this feasible precisely where it is most needed is that the station data doing the correcting need not be dense or continuous. A country's few dozen usable stations, with gaps, are enough to bias-correct a complete field even though they could never fill it directly. Much of that data never reaches global merged products at all: only a handful of synoptic stations from each country are exchanged internationally through WMO's Information System, while national archives hold far more. Ethiopia shares roughly seventeen stations with the world. The national ENACTS record draws on over a hundred reporting every day, and more than five hundred by the end of each year. The same records that were never good enough to initialize a numerical model are good enough for cheap, incremental correction now that using them costs little.

Getting that data into usable shape is treated as an operational goal. Open-source tools assist at each stage: the Automatic Weather Station Data Tool ingests and quality-assures observations from donor-supplied automatic stations that arrive in mutually incompatible formats [85], the Climate Data Tool performs quality control and merging [86], and the Data Sharing Tool releases products on terms the national service sets for itself.

A gold-standard global version would extend this logic operationally. It would recover and quality-control national station archives, blend them against the best available satellite and reanalysis products, and maintain the result through routine workflows as a live reference record. WMO is a natural coordinator, with regional centers perhaps acting as hubs, both because it already operates the sharing infrastructure and standards this depends on, and because the value is realized only when the record is held in common. Building on the national and regional systems that already exist, rather than replacing them, is what would make such a record nationally owned and globally useful at the same time.

Ground truth, data gaps, and the pull toward convenience

Ground-truth data for benchmarking can be thought of in two ways: historical weather data, of any period, to establish local accuracy on hindcasts, and near-real-time data to verify live forecasts. Historical data is more widely available for hindcast evaluation, but both are often acute constraints in LMICs. Where reference data are scarce, benchmarking drifts toward convenience regardless of what is actually needed: models get judged against whatever can be measured easily and not against the decisions that matter, which is itself a force bending AI away from the most exposed populations. Reference data are not monolithic: satellite-only products can fill many gaps but also carry systematic, terrain- and season-dependent biases that station data alone can correct [80]. Blended products that combine satellite and station data, such as ENACTS, are treated as the gold standard for national-scale climate services [81] because they can substantially improve on satellite estimates used on their own. Closing the reference-data gap is what makes decision-relevant benchmarking possible.

Institutions under pressure buy what is easiest to buy. A global foundation model with a dashboard attached is cheaper to procure, faster to deploy, and easier to defend in a budget meeting than a locally calibrated system built on national observations. However, an uncalibrated global model applied to a tropical health decision carries the biases of the data it was trained on, and it may not be appropriate for deployment. Scarce reference data pull evaluation toward what is convenient to measure. Similarly, scarce capacity and a short procurement cycle pull adoption toward what is convenient to buy. Both bend weather intelligence away from the populations with the greatest need. Policy can ensure that this risk is minimized by focusing on procurement standards. Procurement requirements that a deployed system be calibrated against national observations, co-designed with the people who will act on it, and evaluated on a decision metric rather than a global average are specifications a ministry can write and a vendor can meet. Without this kind of deliberate procurement approach, the path of least resistance will decide what models get adopted—wasting public resources and undermining potential public benefits.

The global ground-truth data for LMICs

AI weather models could be trained and tested quickly because the field already had a single record everybody agreed to measure against. A reanalysis takes decades of past observations and reconstructs a complete, gridded picture of the atmosphere at every point and hour, filling the gaps between stations with physics. ECMWF's version, ERA5, is a large part of why AI weather model development moved so fast: models were trained on it and scored against it, so a claim of improvement meant the same thing to everyone [25].

The problem is that ERA5 is weakest in the tropics, where there were fewest observations to reconstruct from [82]. So the reference record is thinnest in the places where the health burden is heaviest, and a model can score well against ERA5 there while still being wrong. Building an observation-based record for those countries would improve both the training data and the standard models are judged against, and it is one of the few investments that improves the forecast and the evaluation at the same time.

There is also data that was never good enough to initialize a physics-based model, such as readings from mobile phones or low-cost stations [65], [66], [67]. Including that data was expensive and marginally useful for numerical weather prediction. For AI models, it costs very little, which changes what is worth collecting.

RECOMMENDATION R-D2

Create a ground-truth weather reference dataset for data-sparse regions

> "There's two tasks here: how do we make these new models, and then how do we actually give people the confidence to take action based on them."
>
> **Niall Robinson**
> **NVIDIA**

CHAPTER 5

Closing the Loop: Decisions, Forecasts, and Feedback

AT A GLANCE

- Demand has to be articulated before a forecast is built, and it takes time and a wide room to cover the range of user needs.
- In Nigeria, a cross-sector exercise found that every agency saw itself as responsible for the warning, which is a different problem from no agency being responsible, and a harder one to see without everyone at the same table.
- A stakeholder-led session to identify weather impacts on health is one step short of a benchmark. Add the threshold that triggers action—noting this should be an impact-based assessment and not just a meteorological threshold—to every decision, the lead time that action needs, and the cost of being wrong, and it becomes a demand-driven forecast specification.
- A system can be built to act efficiently on forecasts and still run on a fixed calendar. Jaipur's heat action plan locates every vulnerable ward but triggers most of its response at the same time each year.
- A forecast can exist and change nothing. In Pune, a dengue model gives six weeks of warning from observed weather, and no health decision has been attached to it, so the value of extending that warning cannot be established.
- Capability and mandate can sit in different buildings. In South Africa, the body with the authority to warn holds no health data, and the body with twenty years of mortality data holds no authority to warn. Neither is refusing to act.
- Across all the cases, a recurring lesson is the transmission gap between meteorological services and health ministries. AI tools can cut that friction where the mandate exists; they cannot create the mandate.

What the cases show

This chapter is about the human-centered design of climate services for health: using AI forecasts to enable new capacities, built around the decisions people need to make. Each case below is a system meeting one of the barriers named at the end of Chapter 3, and in three of the four, the forecast is no longer the one that binds. Demand that was never specified, a forecast that was never built, a decision that was never attached to a forecast, a data foundation that had to be assembled before anything else could start. Together they are an inventory of what a working system needs, taken from the places where each piece is absent.

NIGERIA

Demand, before anything is built

Before a service can be designed around a decision, someone has to ask what decisions need to be made. The Nigeria case study shows the answers when that question is asked properly and across government, and how far short of a forecast specification the answer falls. The largest near-term contributions from AI named in the process were integration, data rescue, and evaluation rather than prediction.

JAIPUR, INDIA

A system waiting for a forecast

Where the decisions, the response tiers, the vulnerability maps, and the delivery channels are already built, the forecast is what is missing, and AI can supply it now. This is the clearest case in this report for building a forecast, and the plan itself specifies which ones it needs, at what resolution, and how decision-makers would act on them.

PUNE, INDIA

A forecast waiting for a decision, and a lead time that does not come from the weather

Dengue's biology supplies six weeks of warning from observed weather alone. A weather forecast extends that only by its own horizon, because the usable warning is set by the shortest predictor lag. The health department asked for the spatial scale of a ward, which is a lot smaller than the 100km grid of the prediction. Where no decision has been attached to a forecast, the question of which improvement to buy cannot be answered at all.

SOUTH AFRICA

The data a health warning needs to function

A health-based warning cannot be set, triggered, or evaluated against health outcomes that were never collected. The South Africa case shows the years of permission-brokering and data assembly that come before any of that is possible. The forecast is absent from this case not because it is unnecessary, but because it is the last component to arrive. Once the foundation exists, it is also, now, the cheapest.

One thing is common across all four. In Nigeria, no agency owned the warning. In Jaipur no institution holds the mandate to reconcile a national warning threshold with a city one. In Pune, the institute that built the model cannot yet issue an advisory. In South Africa, the body with the authority to warn holds no health data. Four cases with four different weak links, and in each one the consistent missing piece is a mandate, and not a weather forecast model. This is discussed more in Chapter 6.

None of the four has a measured health outcome. Jaipur's plan has scaled into a state budget line and into ten more cities without anyone establishing whether it saved a life. South Africa's platform is public, and the decision it was built to support has not changed. Pune's model runs and has never been attached to an action that could be evaluated. Nigeria has no service yet to assess. Four well-resourced, internationally partnered efforts, documented in detail, and not one of them can say whether anyone was better off. This is not a criticism of any of them. It is what happens when evaluation is added at the end and is not designed in at the start, which motivates one of our strongest recommendations in the report.

Renewing the effort of joint production

Case one begins one step earlier than the others. The challenge raised in Chapter 2 returns: for many health applications of forecasting and early warning, the barrier upstream of any forecast is a lack of awareness of the health impacts, or of what forecast products could exist, or an internalized assumption that forecasts cannot do certain things. Left there, the advantages of this weather-forecasting revolution get missed on the basis of what was once possible rather than what is or will be. Development driven by demand takes a step before the forecast: understanding what the demand is, unconstrained by current or past capacities. We begin with one such approach, showing how demand can be articulated before a forecast exists and turned into an actionable set of activities, which differs from the approach of having a forecast and finding problems to set it to.

CASE 1: NIGERIA

Demand articulation before the forecast exists

The decision. Not yet known. We have emphasized the idea that a forecast should be chosen by how well it supports a decision. But how do forecasters know what the decision is? We learn that you have to ask more stakeholders and agencies than you expect, and that what comes back is not a tidy list. It may have critical gaps, lead time among them, because the purpose of the exercise is to identify demand across all lead times at once, before anyone has decided which time range weather intelligence will serve.

Method. The Weather and Climate Information Services (WISER) program, led by the Met Office in the UK on behalf of the UK's Foreign, Commonwealth and Development Office (FCDO), has supported the development of weather and climate information services across Africa since 2015, working with and through nationally and regionally mandated weather and climate producing centers, and advocating a co-production approach which has recently evolved to become "ActionFirst". The approach inverts the usual supply-side ordering. It does not begin with what the weather will do; it begins with what actions can be taken, and works backward to the forecast that action would require.

In coming together to design actionable weather and climate services, stakeholders map what happens in their sectors as a hazard escalates, including what action each escalation demands to protect lives and livelihoods. Impact tables are created to link meteorological hazards with the "so what" of the weather, the underpinning concept of Impact-Based Forecasting approaches, on which ActionFirst builds. In principle, these impact tables, alongside considerations of the likelihood of impacts occurring at the level forecast, provide a link and a trigger for sectoral actions such as Early Action Protocols and standard operating procedures.

Under the WISER Health initiative, a workshop took place in Abuja, Nigeria, in late 2025, in collaboration with NiMet, the Nigerian Meteorological Agency. Stakeholders were invited to draft impact tables, with participants drawn from a diverse range of sectors impacted by weather and climate hazards, including health, emergency management, hydrology, disease control, water, agriculture and livestock, as well as the government entities dealing with statistics, budgets and financing, the underpinning departments needed to support the development of such services and associated action. Community surveys were conducted across fourteen states, covering rural, urban and informal settlements, and a wider workshop followed in Lagos including additional agencies, local government, community-based organizations and local media.

Three findings came out of this exercise.

Nobody owned anything, because everybody owned everything. "When they went through all of what was needed to generate and underpin a warning service, everyone felt they were all responsible for everything, or at least would be under this approach to early warnings," Nyree Pinder of the Met Office said of the result. This is not an obstacle but a principal insight. It points to a fundamental shift in thinking required from the outset, especially in moving toward working under a co-production framework. A forecast can be produced, be skillful, and still have no accountable owner for the warning it becomes, especially if, or when, it is layered with downstream information, or translated by sectors into additional advisories. "Mandate ambiguity" of this kind is invisible at a purely technical scoping, and it is not repairable by a better weather model. It surfaced here because the agencies were in the same room.

Participants asked for forecasts, and stopped one step short of specifying them. The groups named flood forecasts and flood outlooks, seasonal forecasts, daily local weather updates, rainfall information, hydrological data, and health bulletins, alongside vulnerability maps, demographics, health facility and staffing data, and a national dashboard. There was clearly demand for forecasts, but the table identifies the hazard and the product without identifying the threshold that triggers an action, the lead time that action needs, or the cost of being wrong. Those three items are what a decision-relevant benchmark is built from. The table also stops short of why the forecast was needed and how it would be used: what decisions it would support, what actions it would trigger, and who is accountable when it is wrong.

The information the room needed most was often not weather at all. Heat-attributed mortality is named as a gap in Nigeria's Integrated Disease Surveillance and Response Strategy because it is not collected. Without it there is no ground truth against which a heat-health forecast can be benchmarked, and no way to demonstrate afterward that the service worked. WISER's proposed initial fix is a box on a hospital form, where staff can tick the cause of a death or injury, whether flooding, a pothole, or heat. The absence of impact data is a harder constraint on impact-based forecasting than forecast skill is. Around 70% of clinics in Nigeria have no electricity and therefore no refrigeration for cooling supplies, which limits what any warning can achieve.

Running this effort through a meteorological agency alone would leave most of the findings that could benefit health invisible. The route to the household runs through the radio networks; statistical aggregation sits with a bureau the sectors do not feed; the budget sits with economic planning; and half the pathway from rainfall to cholera runs through water and sanitation. The cross-governmental need is only visible when the agencies fill in one table together. This is the case for funding demand articulation as its own deliverable, since exercises of this kind are inexpensive relative to the national investments they steer.

RECOMMENDATION R-E3

Fund demand articulation as a first-class deliverable

Where AI enters, and where it does not yet

Most answers to the question of where AI would help were something other than forecasts.

Integration. Each ministry gathers its own statistics, and the bureau meant to consolidate them is disconnected from those ministries. AI is proposed here as connective tissue across datasets built on incompatible spatial and temporal units.

Data rescue. Historical records are still on paper in many services, and digitized records are often added by hand. Automated scanning with machine quality control is an unglamorous application with direct consequences for how much a data-driven forecast can learn.

Scenario work. Climate risk narratives allow decision-makers to plan for the consequences of adaptation actions in future years, not just the changes in physical climate. This would need sectoral data, and it gives finance and planning ministries something they want, turning data-sharing from an obligation into an incentive.

Evaluation. Attribution is the hard part, and a well-trained model may help quantify effectiveness and avoid high cost. Impact evaluation belongs inside the loop from the beginning, not after the service has run for a decade.

Forecasts. Impact-based forecasting requires combining hazard forecasts with information on vulnerability, exposure, and antecedent conditions to assess risk, which can then be communicated with the support of impact tables and risk matrices to support and trigger action. AI can strengthen this process by integrating diverse datasets and helping forecasters produce more timely impact assessments, especially if they can support the incorporation of more dynamic vulnerability, which is not reflected in standard datasets but rather scraped from situational awareness sources. The challenge is not only technical but institutional. As organizations increasingly develop impact models outside meteorological services, questions remain over who is responsible for producing, accrediting, disseminating, and acting on impact forecasts. Under those conditions, running one's own model is a sovereignty gain and a coordination liability at once. These governance arrangements need to be established alongside technical capability rather than after it.

Where the method reaches its limit

Asking people what they want returns only what they know is available. Urban heat is among the fastest-growing climate-sensitive health burdens, and the participants did not rank it as a risk at all. Perception is shaped by what a hazard has already been seen to do, and by what forecast information has ever been offered. Climate change means not only that existing hazards will grow more severe, but that new hazards may arise. Where a forecast-to-action pipeline was never imagined, asking is necessary, but not sufficient. The exercise needs a step that puts unfamiliar capability in front of the room, or it will keep returning the hazards the room already knows about.

What this case provides

Case one provides a method as well as the information the method drew out. Each finding is something a forecast-first project would have discovered only after the money was spent: that no agency owned the warning, that the impact data needed to verify a heat service did not exist, that most clinics could not refrigerate, and that the agencies responsible could not see one of their largest hazards. What the exercise returned was not a specification, but it was the raw material for one. Add the threshold that triggers action, the lead time that action needs, and the cost of being wrong, and it becomes a decision-oriented benchmark. Case one is where the argument of Chapter 4 begins in practical terms.

Stakeholders. UK Met Office; Nigerian Meteorological Agency; Federal Ministry of Health and Social Welfare; National Emergency Management Agency; Nigeria Hydrological Services Agency; Nigeria Centre for Disease Control; National Orientation Agency; Federal Ministry of Water Resources; Federal Ministry of Budget and Economic Planning; National Bureau of Statistics; Ministry of Agriculture; Ministry of Livestock Development; national community-engagement partners in Nigeria and South Africa; UK Foreign, Commonwealth and Development Office; The Rockefeller Foundation.

CASE 2: INDIA

Building the system before the forecast: heat in Jaipur

The decision. In Jaipur, policymakers must decide whether to move a city from routine readiness into active heat response on a given day, and if so, determine how extensive that response must be. To assist in this effort, the Mahila Housing Trust built the Heat Action Plan with Jaipur Nagar Nigam and the Government of Rajasthan. The decisions around heat action are taken across seven departments: a hospital activating heat-stroke management areas and heat-illness surveillance, the water department positioning tankers into informal settlements, the labor department shifting working hours out of the afternoon, and schools, power, transport and tourism each carrying their own version of these decisions. One color-coded alert triggers a response, and it is issued for that given day.

The stakes. Jaipur's recorded maximum is 47.2°C, and the count of days above 40°C has risen across three decades. The status quo for many of these decisions is to use the calendar: the heat-health ward is made ready when the season opens, around the first of April, and held until July or August whether or not a given week warrants it. Vulnerability is targeted in space and not in time. The counterfactual is not silence. India Meteorological Department heatwave warnings reached the city before the plan existed, and standard actions followed them. What the plan added is who acts, where, and how far the response extends. Heat carries almost no biological lag, so every day of warning has to come from the atmosphere.

> "We just targeted the most vulnerable."
>
> **Bijal Brahmbhatt,**
> **Mahila Housing Trust**

Approach. Jaipur's plan does the hard part. Vulnerability is mapped across all one hundred and fifty wards on parameters chosen for this city, among them slum population density, illiteracy, and access to public transport and to water bodies. These vulnerability indicators were adapted locally rather than borrowed from generic urban heat risk models. Population-group maps place children under five, pregnant women, elderly residents and people with chronic conditions onto primary health center catchments: the wards with the most pregnant women are charted so that the first heat-health ward goes where the budget will otherwise reach the fewest people. A four-tier alert framework grades the institutional response, rising from routine facility stocking through community monitoring of vulnerable individuals to the activation of heat-stroke management areas. An interagency framework maps what each of the seven departments does before, during and after the season. Delivery reaches into the informal sector through auto-rickshaw and jewelry worker unions, slum-level groups, citizens' assemblies, text messaging and physical notice boards, because roughly a fifth of households have no mobile access at all.

The plan is also explicit about what heat risk is, stating that ambient temperature is inadequate, and that wet-bulb globe temperature (WBGT), which combines air temperature, humidity, wind and radiation, is what maps to physiological stress. It computes that quantity for Jaipur from reanalysis, and shows the city crossing occupational danger thresholds through April, May and June. The alerts trigger on forecast maximum temperature. The plan names the right variable and triggers on the one that is produced, and the gap between them belongs to forecast supply.

What changed. The system exists, is used, and has scaled beyond the city. Mahila Housing Trust's cooling station, the first in India, began in Jodhpur; the Government of Rajasthan has since budgeted cooling stations for all forty-one districts, and the Ministry of Housing and Urban Affairs has selected ten cities and given resources to each to act on plans of this kind. The National Disaster Management Authority has since published guidance on cooling shelters, and other cities are replicating the model. Sixteen years of work turned a local innovation into a line in a state budget. What did not change was the timing of most of the response. Daily alerts move resources on the day–the seasonal decisions, the procurement, the staffing and the training–still open and close on the calendar. Whether any of this improved health outcomes has not been measured yet, and a plan operating at this reach is an unusually good place to do it.

Thresholds and evaluation. No threshold in this plan is derived from a health outcome in Jaipur, and three threshold systems are in use: international occupational wet bulb values, city alert thresholds set from the local temperature record, and national criteria set for the country. The last two are printed together but do not agree. A day at 45°C is a red alert by the city's threshold and an orange alert by the national criterion, and no rule

Figure 5.1 The lead-time ladder, Jaipur

Lead time	Decision	Who acts	Currently triggered by
Seasonal	Ward readiness, protocol training, procurement of fluids and cooling equipment; occupational safety guidelines; tanker contingency plans; power load-management plans and hospital backup; school provisioning; transport and tourist-site preparation	All 7 sectors/agencies: Jaipur Municipal Corporation (JMC) / School Board State Health Department Labour Department Public Health Engineering Department (PHED) / Jaipur Municipal Corporation (JMC) Electricity Distribution Companies (DISCOMs) Rajasthan State Road Transport Corporation (RSRTC) / Jaipur City Transport Services Limited (JCTSL) Tourism & Recreation	The calendar
Two to three weeks	Employer shift patterns and heat-break policy; outreach campaign timing; welfare-board and platform compensation windows; parametric product pricing	Employers, platforms, construction worker welfare boards, Mahila Housing Trust	**Nothing**
Three to seven days	Ward activation and staffing against expected admissions; telephonic monitoring lists for high-risk patients; tanker positioning into informal settlements; power prioritization to critical facilities; school advisories	Hospitals, urban primary health centers, water, power, school board	**Nothing.** A seven-day warning already arrives and does not enter the framework.
One day	Alert issuance; labor advisory to move work out of peak hours; facility stocking of fluids and cooling spaces; community monitoring of vulnerable individuals	State disaster management department; municipal nodal officer; labor; health workers; unions and community groups	The alert framework, on forecast maximum temperature
Same day	Activation of heat-stroke management areas; heat-illness surveillance; emergency response	Health system; 108 ambulance service	The alert framework's top tier
Overnight	None specified	—	Night-time heat appears in no alert tier and in no exposure layer

"We initiated a cooling station, which was the first in India. Now the government of Rajasthan has budgeted for all 41 districts having a cooling station. That's the first success."

Bijal Brahmbhatt,
Mahila Housing Trust

states which governs on the day. The National Disaster Management Authority asks cities to set their own thresholds and recommends deriving them from ten years of daily all-cause mortality, falling back to temperature percentiles where that record does not exist. Jaipur had no mortality record and used percentiles. The delegation is sound, and the gap it leaves is operational: a national service defines categories that hold across a subcontinent, a city needs categories that hold in Jaipur, and reconciling the two on a given day has no holder.

The ladder in Figure 5.1 sets out where the plan already acts and where it could. The empty rungs are the opportunity.

The plan's own analysis specifies a forecast need that is not produced. WBGT is already computed for Jaipur from reanalysis by a standard method, and computing the same quantity from forecast fields converts a description of the city's climate into a trigger for its alerts. What AI weather changes here is price and repetition: running several global models, correcting them against local observations including the wet-bulb measurements women's groups already collect on fixed routes, and regenerating at city resolution often enough to be operational. The three-to-seven-day rung needs no new forecasting at all to fill it. A seven-day warning already arrives, and aligning the Heat Early Warning System to it is the cheapest gain available in the plan. Within three to five years, urban-form-aware downscaling may refine forecast specificity. Nighttime heat at kilometer scale may fill the empty overnight rung, because a known health mechanism drives the demand. Further out, generative approaches to compound events, heat arriving with air pollution or over a district that has just flooded, would allow scenario planning that is currently impossible from a record that does not yet contain the event, as well as provide a concrete precursor worth monitoring and a trigger to attach to it.

What a better forecast will not fix. The land allocated for a cooling shelter will still be allocated politically, and it will often not be in the hottest ward. Front-line health centers will still have no refrigeration for ice packs. The warning will still not reach the last-mile health workforce until someone resolves which government employs it. A fifth of households will still have no phone, a heat death will still be recorded as a heart attack, and a worker paid by the day will still go out into the heat, because the wage is the binding constraint, and not the information. The system's own fragility sits in the same category: the citywide alert depends on a municipal heat officer for whom this is an unpaid additional duty, and the warning text was drafted by the non-governmental partner during the plan's first years, before the municipality takes it on. None of this is visible from a forecast. It surfaced because the plan was built with the residents who carry the risk. Pune, which follows, has a forecast and no decision attached to it. Jaipur is the exact reverse, a decision architecture with no forecast underneath it yet, and the pairing is why Chapter 6 discusses institutions rather than models.

RECOMMENDATION R-S3

Fold AI weather for early warning into existing heat action plans, with a drafting tool for low-capacity cities

CASE 3: INDIA

Dengue forecasting in Pune: six weeks of warning from observed weather

The decision. The potential decisions that can be informed by this forecast are clear: when a municipal health department mobilizes ward-level larval source reduction, when it pre-positions beds and fluids ahead of the surge that begins as the monsoon progresses in August and peaks in November, when a state releases the vector-control budget. Each has a different owner, a different useful lead time, and a different tolerance for a false alarm. None has been attached to this forecast yet, but a flexible forecast could be built around thresholds for each action [87].

The stakes. India represents roughly a third of the global dengue burden. Pune, a district of ten million in Maharashtra, recorded 81 dengue deaths in the worst year between 2004 and 2015, and more than eight in ten of its dengue deaths fall after the monsoon breaks in June. Without a forward warning, the operative response is vector control after cases appear. That is the counterfactual, and it is not inaction. It is an action scheduled on a calendar instead of triggered by a signal, which is also why a false alarm here is relatively low cost: a spraying round that would have happened anyway.

Where the lead time comes from. Biology sets the length of the window, and observed weather supplies the signal inside it. The progression from monsoon rains and the resulting standing water available for breeding through to a confirmed dengue death can take up to around 70 days. This period spans the mosquito life cycle, incubation of the virus inside the vector, incubation inside the human host, and the interval between hospital admission and death. That interval is the resource any dengue warning relies on, and a weather forecast would add to it.

The approach. Data on dengue cases from Pune Municipal Corporation and India Meteorological Department's weather data is used with a machine learning method to predict dengue mortality from mean temperature twenty weeks earlier, rainfall eleven weeks earlier, and relative humidity six weeks earlier. This is a health forecast built on observed weather. It needs observations, and not necessarily weather forecasts. The usable horizon is set by the shortest lag rather than the longest, and the model leaves part of the biological lead time unused. A weather forecast could extend that horizon by days or weeks, and the value of the additional lead time will depend on the decision being supported. Whether the observation-based forecast is already all the authorities need for dengue control is an open question. Extending it further could add more, and that turns on there being a decision that needs support at those longer lead times. Nobody has named one, so nobody has had to answer.

Where AI enters, and where it does not. A better weather forecast may add very little. The window already exists, set by the mosquito and the virus, and observed weather already fills it, so lengthening the atmospheric horizon is not the constraint. That does not put AI weather outside the problem. It moves what AI weather contributes, and it makes the value of any extension conditional on a decision that has not been named. Extending the warning remains a real prospect, and finding out when a longer warning would be worth having is a question for the people making the decisions, not the modelers alone. The model runs on gridded temperature at roughly a hundred kilometers for a single city, while the health department has asked for the much smaller ward level. AI-based downscaling could help improve the targeting and specificity of these forecasts, provided it is validated against local observations and health data. Planning for the future is also essential, since the range and incidence of vector-borne diseases will increase. The coarse future projections of climate model output limit this, but fast and extensive AI weather ensembles run under future climate states could help narrow the gap.

"Climate and health data, combined with AI tools, can give us weeks of advance warning, helping health systems act before dengue cases begin to rise."

Roxy Mathew Koll,
Indian Institute of Tropical Meteorology

Who is allowed to warn. India has no agency mandated to produce climate-health forecasts, and that absence is more of a constraint on effective weather intelligence for health than the forecast is. The Indian Institute of Tropical Meteorology (IITM) built the model on individual initiative and, after spending four to five years securing permission to use the mortality data, IITM still cannot issue an advisory. "We do not have the mandate to give forecasts," Koll says. Meanwhile, nonprofits and startups take state contracts to build such systems with data skills but without the inclusion of climate or health expertise. The WMO-WHO joint program has now funded a Climate and Health Desk for South Asia at IITM, the first body positioned to support the development of standards for what an effective climate-health forecast is in the country. Standard-setting authority to accompany the funding would begin to address this mandate issue.

Stakeholders. Indian Institute of Tropical Meteorology; Pune Municipal Corporation Health Department; India Meteorological Department; National Institute of Virology, Pune; Savitribai Phule Pune University; WMO-WHO Joint Office (Climate and Health Desk for South Asia).

CASE 4: SOUTH AFRICA

Assembling the data foundation for heat and air-quality warning

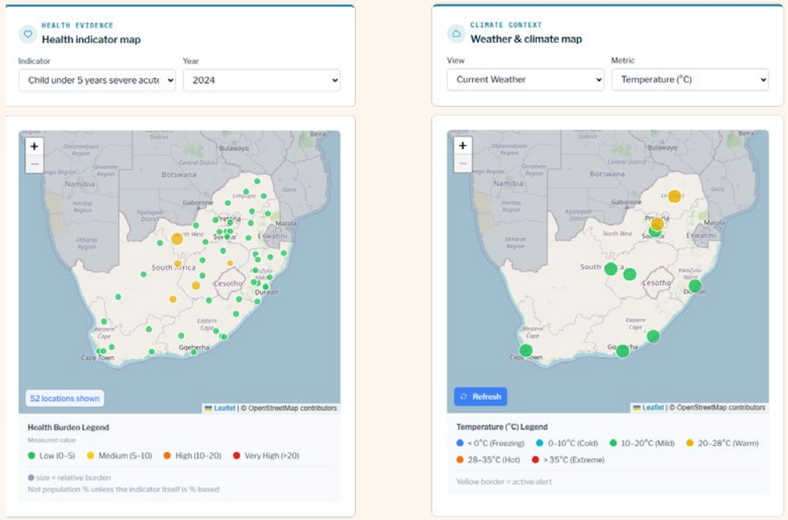
The decision. A clinic, or a household with a pregnant woman or an asthmatic child, may need to prepare for a dangerous heat or air-pollution day, one to seven days ahead. They may need warnings about diarrheal disease or other complications that arise in high heat. That decision is currently unsupported. The South African Weather Service (SAWS) issues warnings on meteorological percentiles, but no health-based warning currently exists. The South African Medical Research Council (SAMRC) is building the data foundation such a warning would require, and the forecast is the last component to arrive.

The stakes and the counterfactual. Without incorporating the information that the health sector needs, a warning activates on the hottest days above a percentile threshold rather than on the days when mortality actually rises. SAMRC derived provincial mortality-based heat thresholds from twenty years of mortality data under a World Bank-funded project, and SAWS has not yet adopted them, retaining a meteorological definition. The counterfactual here is not the absence of a warning system, but a warning system operating on a variable defined by meteorology.

Driver and lag. Heat and air pollution act with near-zero biological lag, which places the usable decision window inside the weather forecast horizon. Diarrheal disease in children under five follows rainfall and flooding at a lag of days to weeks, and that lagged signal is visible in the platform's own overlay of rainfall against case counts.

Approach. The platform assembles, from separate custodians under separate permissions: District Health Information System records from the National Department of Health, aggregated and anonymized, covering diarrheal, respiratory, heat-related, vector-borne and injury outcomes; the met service's register of extreme events; historical climate indicators extracted from a university archive; real-time weather from a public API; roughly 9.5 million air-quality records from 110 monitoring stations; and Ideal Clinic Status rankings, repurposed as a proxy for health-system resilience. Coverage is national, across nine provinces and fifty-two district municipalities. Access to the health data was brokered by a named champion in the National Department of Health's (NDH) Environmental Health Directorate, under an agreement restricting use to this purpose. The forward-looking capacity is not yet a forecast but an exposure-response function, built for heat province by province and for air pollution. A forecast, when the health-demand-driven forecast capacity exists, can use the exposure-response function to tailor triggers to estimated health needs. SAWS does not need the health data. It needs the exposure-response functions health researchers derive from that data. Moving a function is a smaller act than moving a record, and it is easier to govern.

Figure 5.2 SAMRC Climate & Health Surveillance Platform



What changed. The system exists and is public. The decision has not changed and the outcome has not moved, but that is because the platform is creating a foundation for greater capacity to be layered on top. There is no forecasting layered on top of that for now. The health-based awareness messages intended for clinic waiting rooms are designed but not live. This platform sits at the stage before benchmarking, and develops the necessary richness of health data to set empirically derived thresholds for local vulnerable populations. The data foundation now exists, and building it took years of institutional work, paving the way for a health-decision-first system to be created.

That foundation is what made the next step reachable. SAMRC and SAWS are now working together toward a health-based warning, and the work that made the collaboration possible is the assembly that came before it.

The mandate question remains open underneath the collaboration. The health data is owned by the NDH, tax-funded and government-owned. The authority to produce warnings sits with SAWS. The platform is owned and run by a research council holding neither mandate. Two organizations have found a way to work around that, and nothing yet assigns the warning to anyone. Chapter 6 is about what would.

Limitations. Four constraints are recurrent in this case, and none of them concerns the availability or absence of forecasts. SAWS charges for its data, including to another government entity, with no formal exemption available. The word "alert" is reserved for the mandated warning authority, so the platform's own alerts page cannot provide health-based warnings. Once the assembled data proved valuable, a commercial approach exposed an unresolved question about who may derive value from government-owned health records. And the institutional split runs inside SAWS as well as between agencies: the working-level team supports a health-based metric while the decision-making level does not. Once these issues are resolved, layering on improved forecasts built with AI is the cheapest step that remains.

Stakeholders. South African Medical Research Council (lead); National Department of Health, Environmental Health Directorate and District Health Information System; South African Weather Service; University of Pretoria; CICERO; Planet4Health (EU); World Bank; The Rockefeller Foundation; University of East Anglia.

What the cases teach

The lessons below are drawn from four cases in three countries and institutions, and several are specific to the case that produced them. Where a lesson is attached to one setting, it is named as such.

Ask before you build, and expect the answer to be untidy. Demand does not arrive pre-sorted by lead time or by hazard. (Nigeria)

Collaboration and co-production are key in developing weather and climate information services that support action, and for many this is a new way of working, so it will be challenging at the outset. (Nigeria)

Impact-based forecasting is less a forecast modeling challenge than a data integration one. What matters is assembling the data that allow an assessment of the potential level of risk posed by an event, which can then enable actions taken according to individual or sectoral risk appetite, and the cost of not acting. (Nigeria)

The room has to be wider than weather and health. Mandate ambiguity, delivery platforms, missing data, and lack of awareness of key impacts were each visible only from outside the two obvious agencies, and agreeing who runs, accredits and disseminates what has to happen before local model capability arrives, not after. (Nigeria)

Asking what forecasts are needed returns demand as currently perceived, based on what is available. Demand articulation needs a step that shows people capability they have not seen. (Nigeria)

A plan triggers on the variable that is forecast, not the variable that matters. Naming the right one in the analysis does not make it available at the moment of decision. (Jaipur)

Spatial precision and temporal precision are separate achievements. A system can locate every vulnerable ward and still run much of its response on the calendar rather than on a forecast. (Jaipur)

Extra lead time being available does not mean that lead time is used. Before investing in a longer forecast, ask what a decision would do differently if the warning came earlier. (Jaipur, Pune)

Thresholds drawn from the temperature record are not thresholds drawn from harm. A percentile describes a climate; a trigger makes a claim about health. Where the mortality record is missing, the percentile is what is left, and national categories and city categories can both be right and still disagree. The gap between them is a tailoring mandate that often has no holder. (Jaipur)

Lead time comes from disease biology and lagged climate signals together. The shortest predictor lag sets how much warning the system can give. (Pune)

The value of extra lead time is conditional on a decision that can use it, and it does not follow from a longer forecast on its own. If an extra week or ten days over the current six weeks is worth having, the decision is what will say so. (Pune)

The alternative to a warning is rarely nothing. Where a program already acts on a fixed calendar, a forecast has to beat a routine, which is a higher bar and a cheaper false alarm at the same time. This is an essential component of decision-oriented benchmarking. (Pune, Jaipur)

A hundred kilometers and a ward are not separated by forecast skill. Resolution is a claim about the present, and no improvement in prediction closes a gap in the observational record. (Pune)

A capability and a mandate can sit in different agencies, and creating a model and holding the authority to warn are different permissions. In South Africa, the body with the authority to warn holds no health data, and the body with the health data holds no authority to warn. Neither is refusing to act. (South Africa, Pune)

Thresholds are only as good as the institution willing to adopt them. Twenty years of mortality data produced a health-based trigger, and the warning still fires on a percentile while adoption is negotiated. (South Africa)

A platform can only win the agreements its own authority can reach. SAMRC secured a standing feed of health records from a department inside the same government. It could not secure weather data from a service that charges for it and answers to a different mandate, and no amount of platform-building changes that. (South Africa)

Success surfaces the ownership question, it doesn't settle it. Data too fragmented to be valuable raises no question about who may profit from them; assembled data raise it immediately. (South Africa)

Where the forecast is the last thing missing, it is also the cheapest. Years of permission-brokering, assembly and exposure-response work precede a component that can now be acquired rather than built. (South Africa, Jaipur)

The transmission gap

The mandate gap named at the start of this chapter has a specific consequence once forecasts become plentiful. AI tools, both forecasts and language models, can cut the friction between a meteorological service and a health ministry wherever the institutional will exists, and they cannot create the will. Somewhere in the system, a designated body must be responsible for data-sharing, communication, and maintaining and distributing the information produced, and that accountability becomes more important as this capacity expands. Forecasts will inevitably be wrong some of the time, so a system that issues more of them needs a plan for when the decision support fails, including clear accountability for the consequences. This is one of five lessons from agriculture for closing the last-mile gap [11]. The accountability question runs into governance in Chapter 6: expanding the production of forecast-based health information means someone has to own the cases where it fails and weigh the benefits against the costs.

Impact evaluation and feedback

Formal evaluation of whether a warning system changes health outcomes belongs in the loop from the beginning rather than at the end. Each of these four cases would be fertile ground for generating evidence of health effectiveness. That is the argument in its shortest form. The cases show it directly: once a forecast reaches sufficient skill, it is often not the sole binding constraint, and the value is won or lost in translation, delivery, health-system readiness, and the presence or absence of feedback. Evaluation should not capture health outcomes alone. Service access, engagement, recall and comprehension, and trust and belief-updating all indicate whether a disseminated forecast has changed or will change an outcome. Each is measurable, often more easily than the health outcome itself, and measuring them shows where a service is failing and reveals the channels through which people are able to act.

From project to scale

Almost every case here is a project. Getting from a project to a scaled, sustained service is a distinct challenge, and it is where many good innovations stall. The forecast improving is not what carries a project to scale; the institution, the financing, and the coordination to take a proven innovation national are.

BOX 5.1

The scaling problem, and the mechanism that could address it

Most proven interventions never reach scale. This includes even those with strong evidence of effectiveness. Across agriculture, education, and health, the large majority of development innovations with good evidence behind them stall, and the reason is rarely that the innovation was wrong. Innovation ecosystems in low- and middle-income countries lack financing that fits each stage of the scaling journey, and the enabling conditions around an innovation matter more the closer it gets to national coverage [88]. Large sums flow to countries through multilateral development banks (MDBs), and the technical support to design and implement the components those loans pay for is rarely funded, so essential components may be overlooked. Funders running their own innovation programs often have no route from what those programs prove to what their operational units finance. Governments express the demand, but the ministry that holds that sector-specific demand is rarely the one the bank is talking to. None of this is anyone's failure. It is the predictable result of coordination that sits outside every participant's mandate.

AIM for Scale, the Agricultural Innovation Mechanism for Scale, was built to occupy that role in agriculture. Funded by the International Affairs Office at the UAE Presidential Court and the Gates Foundation, it assembles the financial and technical resources needed to take innovations with rigorous evidence of impact and cost-effectiveness to national scale, convening governments, multilateral development banks, bilateral donors, philanthropy, academia, and civil society around a single design. It works only where country demand and a proven innovation already exist. Where they do, it surveys across government to establish who needs what and who holds which capacity, diagnoses why the pieces are not coming together, facilitates technical assistance, connects government institutions to leading technical and implementation partners, and helps translate evidence into the design of large-scale government programs and the MDB-financed investments that pay for them [89]. Weather services for farmers was its first innovation area, with the goal of helping governments reach a hundred million farmers with improved forecasts by 2030 [90].

In Ethiopia, AIM for Scale has supported the government in introducing AI-enabled weather forecasts by bringing together the Ministry of Agriculture, the national meteorological agency, and technical partners to improve forecast accuracy and deliver timely, actionable information through existing government channels. This initial work is also shaping the design of a broader World Bank-supported agricultural investment of roughly $250 million. It acts as a neutral intermediary between the meteorological service and the ministry that will use the forecast, and it carries political weight through a respected figure in the destination sector: a former agriculture minister talking to the agriculture ministry about an agricultural innovation, rather than a meteorological service asking someone to care about weather. Another gap AIM for Scale identified was the technical capacity to adopt and produce AI forecasts, which is a further area of investment: the AI Weather Forecasting for Agriculture Training Program, launched in 2025 with an initial cohort of five countries, brings together representatives from national meteorological services and ministries of agriculture to strengthen institutional capacity to generate, evaluate, and deliver accurate, actionable, farmer-centered forecasts. Across all of this, AIM for Scale's own money is small: grants in the tens to low hundreds of thousands, shaping operations in the hundreds of millions.

Climate and health has the same structural problems for scaling. The innovations are arriving, the demand is there, and the coordination to take innovations to scale remains a challenge. The gap has been identified independently from the health side: the World Bank is not a health agency, and scaling its health work will require an ecosystem of technical assistance providers to help countries design and execute programs, which does not yet exist [91]. A health version of this mechanism needs a messenger carrying a health mandate. Whether it transfers at all rests on two conditions: a diagnosis of why cooperation between meteorological services and the health sector has historically failed, since a mechanism that does not know the answer will reproduce it, and confirmation that health carries the budget priority with target countries and scaling funders to make the model work. What a health version looks like is left open here, deliberately. Four questions must be answered if it is funded. Who hosts it, and does its neutrality survive that host? How is a governing board composed so that the countries it serves hold decision rights and are not relegated to advisory seats? How does it avoid becoming a channel for its funders' priorities? And is it lean, on the AIM for Scale model of a small coordinating team, or inclusive, with the broader membership that legitimacy in the health sector may require? These are questions for co-design, not gaps in the recommendation.

RECOMMENDATION R-S1

Establish or designate a climate-health scaling mechanism

CHAPTER 6

Governance and Coordination

AT A GLANCE

- Institutions that never had to speak to each other now have to, and nothing in the existing arrangements makes that happen.
- Policy should be written for enduring principles, not for specific technologies, because architectures turn over faster than legislation or regulation can follow. Policymakers should prioritize five principles in this domain: data sovereignty, human dignity, contextual utility, public trust, and institutional accountability.
- Regulation and accountability should be tiered by what a decision puts at stake, regardless of how novel the technology is. A public advisory that next week may be warm is not in the same tier as an instruction that routes a limited stockpile to one district over another. Human-in-the-loop verification can limit errors in high-stakes situations.
- A working weather-intelligence-for-health system needs nine capacities, each with a mandated home. This report discusses the capacities rather than the specific institutions, because the right holder varies too much by country to prescribe.
- Authority and legitimacy are different things. A body can hold the authority to issue a forecast without holding the legitimacy—gained through trust, fairness and openness—that makes people act on it, and as models proliferate, that gap widens.
- Accountability for forecast error should be settled deliberately instead of waiting for the first controversial outcome, and settled in a way that does not make officials afraid to act on probabilistic information. Framed as liability for outcomes, accountability produces defensive non-action, and a system that will not act is worse than the absence of information it replaced.

In most countries, the ability to run a weather intelligence system for health already exists, distributed across institutions with no mandate to use it that way. Capacity for the individual technical pieces can be built through training and investments. Building new institutional capability, on the other hand, is slow and expensive. Instead, we emphasize identifying who is responsible for the capability within an existing set of institutions. Three questions guide our approach to discussing governance in this report: Who is responsible? What are they required to do? And, where does that get written down and enforced?

Roles and mandates: the capacities the system needs

AI forecasting will produce a proliferation of weather models, inside and outside national and regional systems. Rather than ask which institution should do what, we identify the capacities a working system needs, each requiring a mandated home. We identify and explain these 9 capacities in Figure 6.1.

1. **Demand articulation and joint production**.
2. **Data stewardship: weather and observational data.**
3. **Data stewardship: health data.**
4. **Forecast creation or identification.**
5. **Benchmarking and forecast validation.**
6. **Operational forecast production.**
7. **Translation.**
8. **Delivery, integration, and transfer.**
9. **Monitoring, evaluation, and feedback.**

In most countries several of these have no identifiable holder. That is the gap, and naming the capacities is what allows the naming of holders to happen.

RECOMMENDATION R-U4

Produce an institutional responsibility model and stakeholder map for governance of AI-driven climate services for health

Authority and legitimacy

When a forecast is wrong, or when a correct forecast is ignored, who is answerable? If better forecasts arrive across a wider range of actionable timescales, someone has to hold the standing to produce them, someone to disseminate them, and someone to require action on them. Those three need not sit with the same body, and in most countries they do not. Sustained action needs both authority and legitimacy, and they are different things.

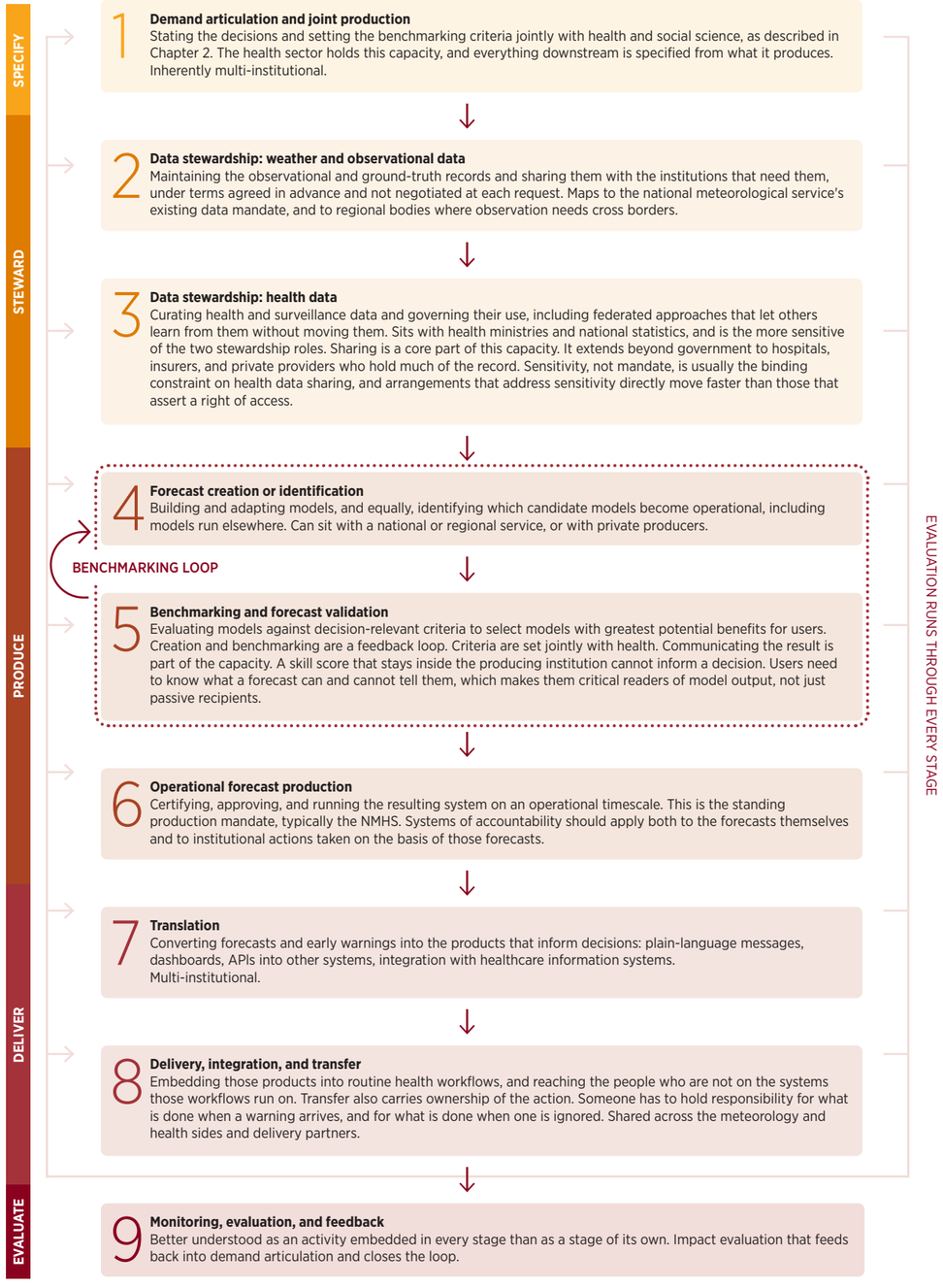
Authority is the express legal, institutional or political standing given to a body to develop and deliver information, and to take or enforce action on it. Where the lines are clear, there is less confusion about what is delivered and to whom, what action is required, and what follows from non-compliance, and less duplication of effort and expense. Where they are unclear, mandate conflict of the kind seen in the Nigeria case emerges. Clear authority is necessary, and it is not costless: strict lines hold institutions in the shape they were given, which slows the cross-sector collaboration decision-first forecasting depends on, and slows adaptation to needs that change faster than mandates do. A weather intelligence service for health therefore has to resolve the following:

- Who holds legal or institutional responsibility for the outcome?
- Who holds the mandate to do the work and to engage the other parties?
- Who holds the technical, scientific, or jurisdictional standing to develop and deliver the product?
- Who is responsible for enforcement, and who bears the consequences of failure?

Legitimacy rests on trust, perceived fairness and openness: how institutions engage with each other, with stakeholders and with the public, how open the process and the communication are, how well the service works, and how trusted the leading body is in that context. Legitimacy is not a fixed attribute. It is held in a setting, with a population, at a time, and where a service has it, people follow it willingly. Establishing it means asking who is leading and whether they are trusted, who is at the table

> "It's not just that people do not want to be the ones who are wrong, but they do not want to be the first to be wrong."
>
> **Titi Akinsanmi, Demistef.ai**

Figure 6.1 The nine capacities a system needs



and how they are engaged over the short and long term, whether stated mission and actions align, and whether different voices and contexts are represented fairly. That last question does not require everyone to hold an equal say. It requires a process that allows genuine engagement.

Authority and legitimacy may not sit with the same agency. Legitimacy may sit with a trusted meteorological service that also holds the authority to produce the forecast. It may equally sit with the community health workers who take that forecast and talk to their neighbors about how to get through the coming heatwave. For at-risk communities, trusted, legitimate sources are often better placed to communicate a health risk and achieve higher rates of protective action. This is especially true if the lines of authority are blurred and multiple sources of conflicting information confuse the audience.

The gap widens as forecasts multiply and more of the chain automates. A health ministry acting on a forecast from a foreign private model, delivered through a platform it does not operate, is exposed to multiple risks and holds explicit duties on none. Authority can also cross borders without anyone conferring it. The US National Hurricane Center serves as WMO's Regional Specialized Meteorological Center for the Atlantic and eastern Pacific, so when an experimental AI model enters its operational workflow, it enters warnings relied on by close to thirty countries, none of which evaluated or mandated it. The arrangement long predates AI and is not the problem. What is new is the speed at which a model can acquire operational uses that nobody decided to give it.

Accountability as a duty of care

Accountability is easier to settle if it is framed as a duty of care rather than as liability for outcomes. A duty of care asks whether an institution met a defined standard of reasonable practice: whether it used a benchmarked model, followed the verification tier set for that decision class, communicated the uncertainty, and acted on the information it had. It does not ask whether the forecast turned out to be right.

That distinction is what makes it safe to act. An official who can show they followed the process is protected when the weather does something else, and an official who cannot show it is not protected by having been lucky. Framed as liability for outcomes, accountability produces defensive non-action, and a system that will not act is worse than the absence of information it replaced. Settled deliberately, this can be a duty. If it is instead settled only at the first controversial or failed outcome, it will be a liability.

Designing policy for enduring principles

AI policy frameworks are being written now, many including health and weather. These should be written for principles. Model architectures, compute paradigms, and software standards turn over on a cycle far faster than legislation can follow, and a policy attached to a specific technology is obsolete before it is enforced. Five principles will hold throughout this cycle:

Data sovereignty, of which ownership is a component. National health and meteorological data should remain under local jurisdictional oversight, which prevents extractive dynamics while remaining compatible with open interoperability standards.

Human dignity, meaning autonomy and privacy. For example, health data is generated by people who did not choose to become AI model training data, and the systems built on those data should respect the choices of those individuals about their involvement.

Contextual utility. Information is useful when it is designed for the uses the people in a specific environment define, instead of what the producer imagines is useful. A great deal of disaster information is produced on the assumption that knowing about a hazard is itself the benefit, when communities frequently understand the protective behaviors they would deploy if they knew in advance they were needed.

Public trust, which is spent faster than it is accumulated.

Institutional accountability, meaning that someone is identifiable in advance as answerable for a decision taken on the basis of a forecast, and that what they are answerable for is a duty of care rather than the outcome of the weather.

Regulation and accountability should both be tiered, with the tier set by what is at stake in the decision. A public advisory that next week may be warm carries a different weight from an instruction that routes a limited medical stockpile, since that involves a potentially consequential tradeoff between one district and another. The first can run automatically. The second has implications for human life and should not be fully automated. The mechanism is a human in the loop, which does not mean every alert is reviewed by a person: national meteorological and health services often have only a handful of staff, and blanket review would turn a safeguard into a bottleneck. Instead, the categories of decision requiring sign-off should be defined in advance, so review is settled before the alert exists instead of negotiated after the fact. The requirement should sit on the health side, where the action is taken and the consequences felt, and the decision classes should be concrete: a rule naming the classes, illustrative examples, and whether sign-off sits at ministry, district or clinic level can be interpreted into practice.

Red teaming before deployment

Tiering raises the question of what a high tier requires. One answer is red teaming: adversarial stress-testing before deployment, at three levels. The technical level asks how the model behaves at the edges, on the anomalies and compound events outside the training record that are frequently the subject of health warnings. The distributional level asks whether errors fall unevenly, misallocating resources away from the sub-populations least able to absorb the loss. The operational level asks whether the alert workflow actually triggers action: whether a warning issued at two on a Sunday morning reaches a district officer with the authority and resources to act. If benchmarking selects the model, impact evaluation asks whether the service changed behavior or outcomes, and regulation governs thresholds and accountability under uncertainty, then red teaming asks how the chosen system fails before anyone depends on it.

> "Do not design your regulations or your policies attached to a technology. Because like the tortoise, you can never catch up with the hare."
>
> **Titi Akinsanmi, Demistef.ai.**

Data governance and sovereignty

AI weather models favor data-rich regions. Training, initializing and benchmarking on local observations change the value of local weather data and improve accuracy at the same time, which makes ownership and access first-order questions. Health data bring constraints weather data does not, privacy chief among them. A near-term mechanism is to share not the data but the learned relationships that inform a decision: publishing a model or projection as a derived product, or using federated approaches without moving raw health records. For well-characterized outcomes such as heat's effect on cardiovascular disease, where the dose-response is established [4], [14], this is sufficient, and existing health-data structures need not be touched. The grander ambition is different: a health foundation model, especially for something as complex as vector-borne disease, would need pooled cross-border data that trusted entities could learn from, which is a substantial challenge requiring dedicated resources.

These workarounds have institutional forms. A sovereign data space is a governed environment in which national health and meteorological data are combined under domestic jurisdiction, with the rules of access and use set locally. A narrower technical version may involve two parties computing on combined data without either taking the other's records away. Both are established practices in other sectors, and in the best-case scenario, neither would require new laws. Some instruments for this exist already: WMO's WIS 2.0 governs meteorological data exchange, FAIR principles govern findability and reuse [92], and the International Health Regulations govern health information sharing between states. Building on these is faster and more durable than writing new frameworks alongside them, and it is the same counsel that applies to regulation more broadly: update on the constant principles rather than overhaul.

RECOMMENDATION R-U3

Health data sharing including federated methods, and the longer-term data-infrastructure fix

Reach

A forecast that cannot reach a person does not fail at the last mile. It fails when it is designed, because the standards, the institutional partnerships, and the delivery platforms were chosen before the forecast existed. Connectivity is distributed inversely to need: 96% of the 2.2 billion people without internet access live in LMICs; internet use is 94% in high-income countries against 23% in LMICs; and mobile broadband remains unaffordable in roughly 60% of LMICs [93]. What follows is that a health warning system designed around a smartphone application reaches almost everyone in the countries facing the least climate-sensitive health burden and a minority of people in the countries facing the most.

Two mechanisms address this inside existing institutions. The first is to strengthen the data-sharing systems that already exist. DHIS2 is used across roughly eighty low- and middle-income countries, and the mandate to share across institutions is already written into how it is meant to work; what has not happened is the actual sharing. Open interfaces make delivering on that easier, and this is the health-side counterpart to meteorology's WIS 2.0. The second is to require that information reach people across multiple platforms. Where a single provider controls the infrastructure through which warnings travel, everyone outside that network is unreachable, and providers do not open their infrastructure voluntarily. Where this has been resolved, regulators required access as a stated principle.

RECOMMENDATION R-S5

Use weather demand to motivate development banks and governments to build the registries, identity and payment systems, and communication networks that health services depend on

Reach is a language question as well as a bandwidth question. A warning delivered in an official national language to a population that conducts its life in another has not reached anyone, and in most of the countries this report concerns, the official language is a minority language. Nigeria's heat and flood warnings have to work in Hausa, Yoruba, and Igbo before they work at all. Translation is not a courtesy applied to a finished product. It is a requirement on the pipeline, cheap to specify at design and expensive to retrofit.

BOX 6.1

Institutions that already exist

Two arrangements already convene the actors a decision-first service needs, and neither was built for health.

Early Warning for All (EW4All) carries a mandate to protect every person on Earth with early warning systems by the end of 2027, organized in four pillars: risk knowledge, observation and forecasting, dissemination, and preparedness to respond. A decision-first approach makes Pillar 1 the place where the decisions are identified and specified, and makes Pillar 4 the place where ownership of the resulting action has to sit. Health is present in the framework but rarely the organizing unit.

Regional Climate Outlook Forums are recurring regional gatherings, convened by WMO with national meteorological services, that produce consensus seasonal outlooks. The idea took shape at a 1996 workshop in Victoria Falls, and the format has persisted for nearly three decades: a training session builds regional capacity, a technical session produces the consensus outlook, and a user session presents it to sectors including health. Co-production was the ambition, and under the constraints of the time, it was a fair approximation of it. Persistence cuts both ways. The forums are already being repurposed, from seasonal outlooks toward subseasonal products, which makes them a plausible host for decisions that need weeks rather than months of warning.

Open interfaces, low-bandwidth standards, and delivery through the channels people actually have, including short message services and community radio, are not degradations of a high-quality message. They are the conditions under which a public warning system is public.

Transparency

Transparency, including transparent benchmarking, is a precondition for institutional trust. The concern that an AI model is a black box is real, but it is not distinctly an AI concern, for reasons Chapter 4 sets out. The governance implication that follows from this is that a body that cannot show how its forecasts were selected and how they have performed has nothing to stand on when one of them is wrong.

BOX 6.2

Where health sits in national data and AI policy

Every country revising a data policy, drafting an AI strategy, or setting rules for automated decision-making is writing the rules that will govern weather intelligence for health, usually without knowing it. These instruments are on short cycles and durable once written.

The recognition is already there. The African Union's Continental AI Strategy [94] names health among its priority sectors, and most national strategies do the same. What is largely absent is the layer below recognition: a named lead agency, a mandate for data exchange between health and meteorological authorities, a budget line, and an accountable owner when an automated warning is wrong. This is the pattern that played out in climate policy one instrument earlier, where health was named in almost every national adaptation plan and operationalized in very few.

Two features of the health sector make the risk worse. It is heavily regulated, so a general-purpose AI rule written without health in mind will either exclude health applications or subject them to verification requirements designed for a different risk profile. And health data is governed by consent and privacy regimes that predate any of this, so a data-sharing provision written for open government data will not reach them.

What a national policy should require. Six requirements that follow from the principles above. Each is written to survive a change of model architecture.

1. **Sovereign data spaces and clean rooms**. National health and meteorological data combinable under domestic jurisdictional oversight, in governed environments where the rules of access, use, and audit are set locally. Specify the environment, not the technology that runs inside it.
2. **Interoperability against open standards**. Health data structured to HL7 FHIR, meteorological data to WIS 2.0, both discoverable under FAIR principles. Alignment of standards belongs in procurement conditions instead of residing only in a strategy document.
3. **Tiered verification by decision class**. Define in advance which categories of decision may run automatically and which require sign-off, and at what level. Weight the requirement toward the health side, where the action is taken and the consequences felt.
4. **Red teaming before deployment**. Adversarial testing at three levels before public rollout: technical behavior at climate edge cases, distributional bias in whom the errors fall on, and operational simulation of whether an alert triggers action on the ground.
5. **Multi-channel and multi-language delivery**. Alerts reaching people through lightweight open interfaces feeding SMS, community radio, and the channels people already use, in the languages they already use. Cross-platform reach should not be purely voluntary. It should be stated in license conditions.
6. **Local calibration as a procurement condition**. Any deployed system calibrated against national observations, co-designed with the people who will act on it, evaluated on a decision metric rather than a global average, and generating evidence of effectiveness for health.

Climate and health in national policy frameworks

Health is mentioned but not operationalized in the policy instruments now being written to govern AI, as it was in climate policy before them [94]. Health is heavily regulated, which slows adoption relative to sectors with lighter oversight, and its connection to meteorological services is thin in most countries, so neither side is well placed to press for the other's inclusion. National adaptation plans make the pattern clear. Across fifty-nine plans [31], every one identifies health as a vulnerable sector. Fewer than half, 42%, name a ministry of health as the lead agency for the health section, and 17% specify a climate-change focal point inside the health ministry. Only two in five estimate the financial costs of their health measures. Recognition is almost universal and unaccompanied by action. On financing, 0.2% of bilateral and multilateral adaptation finance has health as its primary focus [2], and health accounts for under 1% of budgeted adaptation totals where budgets exist at all [33].

As national AI and data policies are being drafted, the concern is not that policymakers will ignore health entirely, but rather that they will see health as too challenging to address in the first generation of AI governance frameworks. As in climate adaptation, the result will be policies that pay lip service to health without taking meaningful action.

RECOMMENDATION R-U2

Health in national AI and data policy frameworks, including data licensing, benefit-sharing, and tiered accountability by decision class

None of this is proposed into empty institutional space. Early Warnings for All offers guidance on how AI weather might be integrated in a health-relevant policy setting (Box 6.1). Regional convening mechanisms are a second hook: the Regional Climate Outlook Forums have brought meteorological services and sectoral users, health among them, into the same room three or four times a year for close to three decades. They could become a natural home for demand articulation, and could carry decision-oriented benchmarking without a new body being created to hold it.

Who needs to be at the table

The list is long, and the convening is unowned. A working service needs the meteorological service, the health ministry and its disease programs, subnational health authorities, the disaster management agency, the data protection authority, the finance ministry that holds the budget line, and whoever actually reaches the affected population, which is often a community health worker or an extension agent rather than an institutional message from a ministry. No single body has a mandate to bring them together, and the absence of a convener is more often the binding constraint than the absence of willingness. The mandate to convene becomes an important role in its own right.

CHAPTER 7

Recommendations, Pathways, and Opportunities

AT A GLANCE

- Recommendations aim to: 1) fund forecasts and systems by whether they change a health decision for the better, not by whether they reproduce the atmosphere; and 2) treat equity as a design requirement, naming where decision-making, budgets, authorship, and long-term control will sit.
- Investment is organized into four types: Discovery, Evidence, Scaling, and Sustainability. Evidence sits before Scaling deliberately, because a scaling funder cannot act without it, which makes evidence generation the unlock rather than a parallel activity.
- Scientific and economic incentives push against the work this field needs most, which makes the case for catalytic capital. Academic publication and commercial rewards pull effort toward the model-building frontier and away from operational science that makes a forecast usable for health: many of the benefits are public goods (high social value, low private return) that private capital will not supply.
- Better forecasts do not by themselves produce action, and there is almost no evidence that warning systems change health outcomes. Confidence in a forecast's reliability produces action, together with accountability settled in advance as a duty of care rather than liability for outcomes.
- Catalytic capital is small and well-placed, unlocking much larger flows over time, including within national budgets and from regional, multinational, and private institutions.
- Almost every effort described in this report is a project. The step that has no owner is the one from a working project to a sustained service, and it is a different problem from the one that got the project working.
- Procurement rules decide what can be bought before anyone decides what is worth building. Where they require a finished product, they exclude the joint specification that decision-driven services depend on.

Two commitments shape how the recommendations below were chosen. The first is to fund forecasts and systems by whether they change a health decision, not by whether they reproduce the atmosphere. Each recommendation is therefore specified against the decision it serves, naming who owns that decision, what authority they hold to act on it, what resources they control, and what would change if the forecast improved. A recommendation that cannot answer those questions is not yet fundable. The second is that equity considerations enter at the beginning rather than arriving as a downstream benefit. Where a recommendation builds a dataset, a benchmark or an institution, it must say where decision-making, budgets, authorship and long-term control will sit, and place them closer to the people carrying the greatest unmet burden. Without that, the agenda improves global forecasting capability and leaves power and infrastructure concentrated where they already are.

Sequence of four investment types

Discovery is the catalytic capital that shapes the direction of AI weather research toward health-decision relevance. Funding for identification of problems and for forecasting competitions belongs in this category.

Evidence is dedicated funding to establish whether forecast innovations actually change actions or health outcomes. It deliberately precedes Scaling: the field should not scale approaches until they have demonstrated improved outcomes. Evidence generation should be built into projects early.

Scaling is development-bank, bilateral, or private capital to roll out what works, unlocked by the economic case and by country-level investment cases. Private and for-profit capital belongs here, at scaling, not earlier, so the agenda is not driven prematurely by profit.

Sustainability is the workforce, institutional capacity, and data infrastructure that grant cycles cannot sustain, and that must be integrated into national budgets and mandates, or sustained by mature market demand for the health services.

Two incentive systems push against AI weather innovation reaching those currently underserved, and together they make the case for catalytic capital.

> "There is a need to demonstrate a convincing rationale and practical pathway for government engagement. Governments appear interested and willing to participate, but they need clarity on how they should engage and what that engagement would involve in practice."
>
> **John Bosco Isunju, Makerere University**

The scientific incentive rewards novelty: a new modeling technique carries academic publication and career returns, while the applied, operational science that makes a forecast usable for a health decision carries almost none. The first will happen without philanthropic investment; the second will not. The economic incentive rewards appropriability, the ability to capture the returns from what someone creates. Many of the benefits of weather intelligence for health are public goods, high in social value and low in private return, so they will not attract private capital. Where both incentives point away from work with high social value, a well-placed grant changes the trajectory.

Catalytic funding will not always be successful. These investments will not deliver where country demand is weak, where the authority to act on a forecast is unclear, or where the resources to respond are absent. They will not deliver where uncertainty is communicated poorly, so that warnings either paralyze or, through repeated false alarms, spend the trust they depend on. They will not deliver where agenda-setting is captured by vendors, where the datasets and benchmarks that anchor evaluation are themselves inequitable, or where a service is scaled before there is evidence that it changes behavior or health outcomes. The recommendations that follow are chosen partly for how directly they address these conditions, and for the existence of safeguards that can minimize those risks.

Different funders suit different points in the sequence. Philanthropic and catalytic capital belong at the front: defining the health decision, building open benchmarks and public-good datasets, funding evaluation, designing equity safeguards, and supporting the early design of institutions before any can pay their own way. Development banks and bilateral funders suit scaling, once there is credible evidence of benefit and genuine country demand. Governments own the long horizon: the budget lines, mandates and institutional homes that carry a service after grant cycles end. The private sector may enter in several ways once the regulatory environment is developed, from early innovators to at-scale service providers, and may be relevant early through interoperability, procurement design, maintenance, deployment and responsible-AI assurance within a governed ecosystem. The question is less when private actors participate than under what rules, with what safeguards, and in whose interest.

RECOMMENDATION R-S2

Unbundle procurement, funding the problem rather than the model

BOX 7.1

If you fund one thing

Each investment type has a different funder and a different single ask. One more could easily have been swapped in: **the subseasonal forecasting-for-health consortium (R-D6)**.

Discovery, for research philanthropy, science funders, and technical foundations already invested in AI weather: build the benchmarking infrastructure, starting with heat (R-D3). Heat is the right entry point because the decisions are well understood, and the infrastructure transfers to other outcomes once built. This needs money and a decision and nothing else, which makes it the fastest thing on this list.

Evidence, for health funders and development-research funders: require every service you fund to specify, before it is built, the health decision it serves and to be evaluated against that decision from its first cycle, and pay for the facility that makes rigorous evaluation possible (R-E3, R-E2). A service that has not stated what decision it supports, who owns that decision, and what a better forecast would change has nothing to be evaluated against. Very little evidence exists showing that warning systems change health outcomes, which is the input scaling capital cannot move without. Evaluation built in at rollout costs a fraction of generating the same evidence afterward, and it can measure behavior change as well as outcomes.

Scaling, for development banks, bilateral donors, and the foundations that work with them: build the institution that carries evidence-backed climate-health services to national scale (R-S1). An AIM for Scale for health. Grants of a few hundred thousand at this point shape loans in the hundreds of millions.

Sustainability, for governments, policy funders, and bilateral technical assistance: get health into the national AI and data policy frameworks being written now (R-U2). These instruments are on short cycles and hard to reopen once written.

Priorities by time horizon

Recommendations are also grouped by when they pay off: about a year, three to five years, and toward ten. The ten-year horizon is defined by what health experts say they need and worked backward, instead of forward from what is technically convenient. The horizon each recommendation sits on is a column in the table below.

A single first ask for each investment type, and the funder each is addressed to, is set out in Box 7.1. Many of these recommendations point to structural interventions in the enabling environment, and will need funders whose appetite runs to the hard enabling investments in addition to piloting innovations. The Climate and Health Coalition is a strong candidate home for a scaling mechanism. Development banks are a natural source of scaling finance—the World Bank in Africa, IDB in Latin America, ADB in Asia—alongside repurposing existing funds such as Universal Service Provision Funds and World Bank Smart Bids. The economic case that unlocks scaling is the benefit-cost range of $3.60 to $68.40 per dollar [95] against the projected health burden. AI forecasting will dramatically lower the cost part of this calculation, and the benefit-cost ratio will increase.

BOX 7.2

Funding the problem, not the model

A common model for both philanthropic and government initiatives that aim to apply novel technologies to persistent social problems involves funding a coalition of actors to collaborate. The impulse to "break down silos" and encourage cross-disciplinary work is not a bad one—it is just too often applied too early in the process, before a problem has been specified clearly enough. In the emerging field of AI weather prediction, this challenge is further compounded by the fact that no single AI weather model (or any weather model) is the single best model: all will have strengths and weaknesses for different geographies, weather phenomena, and ultimately decisions. That means that if a project team includes a single AI company or selects a single model at the outset of the project, the model may not be best suited for addressing the problem at hand.

An alternative approach is possible, and would use multiple funding approaches over time. First, catalytic philanthropic investment can be focused on problem definition. As the Nigeria case study illustrates (Case 1: Nigeria), this is not always an easy or straightforward step, and it can be made even more challenging as climate change increases the potential for jurisdictions to undergo novel or complex weather events they have never before experienced. Philanthropic funders could provide relatively modest grants to research teams to interview decision-makers at varying levels—from individuals to large organizations—to develop a more refined understanding of what kinds of forecasts are actually useful for influencing decisions for given geographies, hazards, and health impacts. These findings can then be used to run more structured multi-stakeholder sessions and build buy-in for necessary changes to the policy environment, as needed.

After the problem has been clearly defined, a later stage of funding can put different AI models in direct competition with one another to determine which is most effective at delivering a forecast that is actually responsive to the problem. As noted elsewhere, the most useful model may or may not be the same as the model that delivers the most technically accurate or specific rendering of the atmosphere – but the question of which model, or which combination of models, will be best is difficult to predict. In addition to better serving particular problems or projects, a research and funding approach that puts models more directly in competition with each other may also help reinforce the importance of centering the health decision, and not just forecast accuracy, in benchmarking and further model development.

The Recommendations

Throughout the process leading to this report, the primary selection criterion was to fund forecasts and systems by whether they change a health decision, not by whether they reproduce the atmosphere.
Twenty-three recommendations across the four investment types. Each is described in full in the Appendix. The chapter each arises in is given so the argument behind it can be found.

INVESTMENT TYPE		RECOMMENDATION	NEAR TERM: 1 YEAR	2 YEARS	3 YEARS	MEDIUM TERM: 4 YEARS	5 YEARS	LONG TERM: 6-10 YEARS
DISCOVERY the catalytic capital that shapes the direction of AI weather research toward health-decision relevance. Funding for identification of problems and for forecasting competitions belongs in this category.	**R-D1**	**Health-and-impacts gold-standard dataset** · *Ch2 · 1-3y*	■	■	■			
	R-D2	**Ground-truth weather reference dataset for data-sparse regions** · *Ch4 · 3-5y*	■	■	■	■	■	
	R-D3	**Decision-oriented benchmarking of forecasts for health** · *Ch4 · 1y*	■					
	R-D4	**Climate-health model intercomparison project (CH-MIP)** · *Ch2 · 3-5y*	■	■	■	■	■	
	R-D5	**Shared tooling for adapting and evaluating forecast-improvement techniques across the resource spectrum, benchmarked against health-relevant extremes** · *Ch3 · 1-3y*	■					
	R-D6	**Subseasonal-to-seasonal forecasting-for-health consortium** · *Ch2, Ch3 · 3-5y*	■	■	■	■	■	
	R-D7	**Training from observations, including satellites, for data-sparse regions** · *Ch3 · 3-5y*	■	■	■	■	■	
	R-D8	**Novel and non-traditional data sources for localization and downscaling** · *Ch3 · 1-3y*	■					
	R-D9	**Generative scenario models conditioned on health-relevant compound events** · *Ch3 · 3-5y*	■	■	■	■	■	
	R-D10	**Regional model-training and localization centers** · *Ch3 · 3-5y*	■	■	■	■	■	
	R-D11	**Open-source or benchmarkable standard for AI weather models** · *Ch4 · 1y*	■					
EVIDENCE dedicated funding to establish whether forecast innovations actually change actions or health outcomes.	**R-E1**	**Impact evaluation of health benefits from forecasts, including assessment of communication and the decision boundary** · *Ch2 · 1-5y*	■	■	■	■		
	R-E2	**An evaluation and match-making facility, pairing researchers with implementers and building evaluation into existing programs** · *Ch2 · 1y*	■					
	R-E3	**Demand articulation as a funded, first-class deliverable, with research on protective behaviors at newly available lead times** · *Ch2, Ch5 · 1-3y*	■	■	■	■		
SCALING is development-bank, bilateral, or private capital to roll out what works, unlocked by the economic case and by country-level investment cases.	**R-S1**	**Climate-health scaling mechanism** · *Ch5 · 3-5y*	■	■	■	■	■	
	R-S2	**Unbundled procurement, funding the problem rather than the model** · *Ch7 · 1-3y*	■	■	■	■		
	R-S3	**AI weather for early warning folded into existing heat action plans** · *Ch5 · 1y*	■					
	R-S4	**Monitoring and surveillance integration for climate-sensitive disease** · *Ch2 · 1-3y*	■	■	■	■		
	R-S5	**Weather as a use case for digital public infrastructure, using weather demand to motivate patient and farmer registries and shared identity and payment rails** · *Ch6 · 3-5y*	■	■	■	■	■	
SUSTAINABILITY is the workforce, institutional capacity, and data infrastructure that grant cycles cannot sustain, funded by national budgets or market demand.	**R-U1**	**Capacity realignment: training forums, summer schools, embedded fellowships** · *Ch3 · 3-10y*	■	■	■	■	■	■
	R-U2	**Health in national AI and data policy frameworks, including data licensing, benefit-sharing, and tiered accountability by decision class** · *Ch6 · 1-3y*	■	■	■	■		
	R-U3	**Health data sharing, including federated methods, and the longer-term data-infrastructure fix** · *Ch6 · 3-10y*	■	■	■	■	■	■
	R-U4	**Institutional responsibility model and stakeholder map for governance of AI-driven climate services for health** · *Ch6 · 1-3y*	■	■	■	■		

Conclusions

A second revolution in weather forecasting is underway, and it will reach public health whether or not the health field shapes it. For 70 years, the capacity to produce a good forecast sat with a handful of well-resourced centers, and the billions most exposed to climate-related health risk were largely left out. That concentration was not a choice anyone made. It followed three constraints: the cost of running a model, the time it took to change one, and the difficulty of tailoring output to a place. AI has lifted all three, and quickly.

For many of the populations this report concerns, the forecast has been a real barrier: too coarse, too late, or built around a variable that does not track what makes people ill. Relieving that is a significant achievement, and it should not be undersold. It is also not the same as improving climate services. Low forecast quality was a convenient answer to why a service underdelivered, and as it stops being the binding constraint it stops being the answer. What it was holding in place comes into view: decisions that were never specified, metrics that were never matched to the health effect, lead times that were never aligned to what a health system does, ground truth that was never collected, institutions that never had to speak to each other, and health benefits that were rarely measured.

None of these challenges is new, and none of them is easy. What has changed is that they are now worth facing directly. Each was intractable for the same reason: solving one alone, while the forecast remained low quality, delivered nothing. That is why they persisted together, and it is why the return on addressing them has risen even though the difficulty has not fallen. It is also why the opportunity is largest exactly where the record is worst. Where every link in the chain is weak at once, incremental repair was pointless, and coordinated investment is now worth more than the sum of its parts.

The practical consequence is a change in where a service begins. Producing climate information for health has usually started from what could be produced, and it can now start from what has to be decided. The amount of advance warning a decision needs is set by the biology of the health impact and the logistics of the health system, not by what forecasting produces most readily, and that is what should determine which forecast is built. How models are evaluated is what decides which models are developed next, so the tests, evaluations, and benchmarks a field uses are not a technical detail but the steering mechanism for the whole enterprise. And better forecasts do not by themselves produce action. Confidence in their reliability does, together with accountability for errors settled in advance as a duty of care rather than as liability for outcomes.

Very little of this happens as a by-product of pursuing better representations of the atmosphere. It happens if the institutions, the rules and the resources now being set are written with health decisions in view, and if equity considerations enter at the beginning of forecast production and not just as a downstream benefit of having produced the right forecast. Those instruments are being drafted now, on short cycles, and they are durable once written. The risk is not that they ignore health. It is that health is named without any attached action or responsibility, as it was in climate policy one instrument earlier.

The technology to change what forecasting can do for health already exists. Whether it is built around the decisions that protect the people who need it most is a choice, and it is being made now by default wherever it is not made deliberately.

Bibliography

[1] G. Cissé et al., "Health, Wellbeing, and the Changing Structure of Communities.," in *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1st ed., Cambridge University Press, 2022. https://doi.org/10.1017/9781009325844.

[2] World Meteorological Organization, *2023 State of Climate Services: Health*. WMO, 2023. library. https://wmo.int/records/item/68500-2023-state-of-climate-services-health.

[3] M. Linsenmeier and J. G. Shrader, "Global inequalities in weather forecasts," *Nat. Commun.*, Sep. 2026, https://doi.org/10.1038/s41467-026-77630-w

[4] T. Carleton et al. "Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits" *Q. J.Econ.*, vol. 137, no. 4, pp. 2037-2105, 2022. https://doi.org/10.1093/qje/qjac020

[5] M. C. Thomson and S. J. Mason, Eds., Climate information for public health action, 1st ed. in Routledge studies in environment and health. London: Routledge, 2019.

[6] R. B. Alley, K. A. Emanuel, and F. Zhang, "Advances in weather prediction," *Science*, vol. 363, no. 6425, pp. 342–344, Jan. 2019, https://doi.org/10.1126/science.aav7274.

[7] D. W. Cash, J. C. Borck, and A. G. Patt, "Countering the Loading-Dock Approach to Linking Science and Decision Making: Comparative Analysis of El Niño/Southern Oscillation (ENSO) Forecasting Systems," *Sci. Technol. Hum. Values*, vol. 31, no. 4, pp. 465–494, Jul. 2006, https://doi.org/10.1177/0162243906287547.

[8] J. W. Hansen, "Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges," *Agric. Syst.*, vol. 74, no. 3, pp. 309–330, Dec. 2002, https://doi.org/10.1016/S0308-521X(02)00043-4.

[9] M. C. Lemos, C. J. Kirchhoff, and V. Ramprasad, "Narrowing the climate information usability gap," *Nat. Clim. Change*, vol. 2, no. 11, pp. 789–794, Nov. 2012, https://doi.org/10.1038/nclimate1614.

[10] K. Findlater, S. Webber, M. Kandlikar, and S. Donner, "Climate services promise better decisions but mainly focus on better data," *Nat. Clim. Change*, vol. 11, no. 9, pp. 731–737, Sep. 2021, https://doi.org/10.1038/s41558-021-01125-3.

[11] K. Baylis et al., "Five Lessons for Closing the Last Mile: How to Make Climate Decision Support Actionable," *Earths Future*, vol. 13, no. 8, p. e2024EF005799, Aug. 2025, https://doi.org/10.1029/2024EF005799.

[12] E. Coughlan De Perez et al., "Epidemiological versus meteorological forecasts: Best practice for linking models to policymaking," *Int. J. Forecast.*, vol. 38, no. 2, pp. 521–526, Apr. 2022, https://doi.org/10.1016/j.ijforecast.2021.08.003.

[13] C. Saulo, "WMO 75th anniversary event for Genève International [opening remarks]," Nov. 2025, Accessed: Sep. 02, 2026. [Online]. Available: https://wmo.int/content/wmo-75th-anniversary-event-geneve-international

[14] A. Gasparrini et al., "Mortality risk attributable to high and low ambient temperature: a multicountry observational study," *The Lancet*, vol. 386, no. 9991, pp. 369–375, Jul. 2015, https://doi.org/10.1016/S0140-6736(14)62114-0.

[15] T. Carleton, E. Duflo, B. K. Jack, and G. Zappala, "Adaptation to climate change," in *Handbook of the Economics of Climate Change*, vol. 1, Elsevier, 2024, pp. 143–248. https://doi.org/10.1016/bs.hesecc.2024.10.001.

[16] F. Rabier, A. Brown, M. Chantry, and F. Pappenberger, "Weather forecasting in a changing climate: the rise of AI and Machine learning?," *J. Eur. Meteorol. Soc.*, vol. 4, p. 100040, Jun. 2026, https://doi.org/10.1016/j.jemets.2026.100040.

[17] S. Lang et al., "AIFS -- ECMWF's data-driven forecasting system," Aug. 07, 2024, *arXiv*, https://doi.org/10.48550/arXiv.2406.01465.

[18] A. H. Murphy, "What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting," *Weather Forecast.*, vol. 8, no. 2, pp. 281–293, Jun. 1993, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

[19] J. E. Bell et al., "Changes in extreme events and the potential impacts on human health," *J. Air Waste Manag. Assoc.*, vol. 68, no. 4, pp. 265–287, Apr. 2018, https://doi.org/10.1080/10962247.2017.1401017.

[20] A. J. McMichael, "Global environmental change and human population health: a conceptual and scientific challenge for epidemiology," *Int. J. Epidemiol.*, vol. 22, no. 1, pp. 1–8, Feb. 1993, https://doi.org/10.1093/ije/22.1.1.

[21] "NOAA's Eyes in the Sky - After Five Decades of Weather Forecasting with Environmental Satellites, What Do Future Satellites Promise for Meteorologists and Society?," *World Meteorological Organization*. Accessed: Aug. 03, 2026. [Online]. Available: https://wmo.int/resources/bulletin/vol-62-1-2013/noaas-eyes-sky-after-five-decades-of-weather-forecasting-environmental-satellites-what-do-future

[22] M. Ventures, "2025: The State of AI in Healthcare," *Menlo Ventures*. Accessed: Aug. 03, 2026. [Online]. Available: https://menlovc.com/perspective/2025-the-state-of-ai-in-healthcare/

[23] G. P. Cressman, "Numerical Weather Prediction in Daily Use: The tenth anniversary of calculations for daily weather forecasts finds forecasts more accurate and useful." *Science*, vol. 148, no. 3668, pp. 319–327, Apr. 1965, https://doi.org/10.1126/science.148.3668.319.

[24] J. Pathak et al., "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators," Feb. 22, 2022, *arXiv*, https://doi.org/10.48550/arXiv.2202.11214.

[25] "The ERA5 global reanalysis - Hersbach - 2020 - Quarterly Journal of the Royal Meteorological Society - Wiley Online Library." Accessed: Jul. 31, 2026. [Online]. Available: https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3803

[26] O. Jay and F. Jahan, "The science of climate heat risks has a metric problem," *Commun. Earth Environ.*, vol. 7, no. 1, p. 595, Jul. 2026, https://doi.org/10.1038/s43247-026-03804-5.

[27] J. Shumake-Guillemot and L. Fernandez-Montoya, "Climate Services for Health," 2018.

[28] E. Coughlan de Perez et al., "Landscape of Anticipatory Action for Health in a Changing Climate," 2025.

[29] N. Angrist, A. Beatty, C. Cullen, and T. M. Nkwane, "Iterative A/B Testing for Social Impact: Rigorous, Rapid, Regular," 2026, https://doi.org/10.48558/1K00-YD44.

[30] T. Tiggeloven et al., "The role of artificial intelligence for early warning systems: Status, applicability, guardrails, and ways forward," *iScience*, vol. 28, no. 11, Nov. 2025, https://doi.org/10.1016/j.isci.2025.113689.

[31] World Health Organization, "Health at the heart of national adaptation planning: a global review of national adaptation plans and health national adaptation plans," Geneva, 2025. [Online]. Available: https://iris.who.int/server/api/core/bitstreams/a0dbe884-2467-4f2d-975c-c2eda77b7dd2/content

[32] K. L. Ebi et al., "Burning embers: synthesis of the health risks of climate change," *Environ. Res. Lett.*, vol. 16, no. 4, p. 044042, Apr. 2021, https://doi.org/10.1088/1748-9326/abeadd.

[33] World Health Organization, "Health in national adaptation plans: review," 2021.

[34] Q. Zhao et al., "Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: a three-stage modelling study," *Lancet Planet. Health*, vol. 5, no. 7, pp. e415–e425, Jul. 2021, https://doi.org/10.1016/S2542-5196(21)00081-4.

[35] A. Osgood-Zimmerman et al., "Mapping child growth failure in Africa between 2000 and 2015," *Nature*, vol. 555, no. 7694, pp. 41–47, Mar. 2018, https://doi.org/10.1038/nature25760.

[36] N. Rieke et al., "The future of digital health with federated learning," *Npj Digit. Med.*, vol. 3, no. 1, p. 119, Sep. 2020, https://doi.org/10.1038/s41746-020-00323-1.

[37] S. Rasp et al., "WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models," *J. Adv. Model. Earth Syst.*, vol. 16, no. 6, p. e2023MS004019, Jun. 2024, https://doi.org/10.1029/2023MS004019.

[38] X. Basagaña and J. Ballester, "Unbiased temperature-related mortality estimates using weekly and monthly health data: a new method for environmental epidemiology and climate impact studies," *Lancet Planet.* Health, vol. 8, no. 10, pp. e766–e777, Oct. 2024, https://doi.org/10.1016/S2542-5196(24)00212-2.

[39] C. Rosenzweig et al., "The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies," *Agric. For. Meteorol.*, vol. 170, pp. 166–182, Mar. 2013, https://doi.org/10.1016/j.agrformet.2012.09.011.

[40] S. Asseng et al., "Uncertainty in simulating wheat yields under climate change," *Nat. Clim. Change*, vol. 3, no. 9, pp. 827–832, Sep. 2013, https://doi.org/10.1038/nclimate1916.

[41] K. A. M. Gaythorpe et al., "Estimating the impact of vaccination: lessons learned in the first phase of the Vaccine Impact Modelling Consortium," *Gates Open Res.*, vol. 8, p. 97, Sep. 2024, https://doi.org/10.12688/gatesopenres.15556.1.

[42] N. G. Reich et al., "Collaborative Hubs: Making the Most of Predictive Epidemic Modeling," *Am. J. Public Health*, vol. 112, no. 6, pp. 839–842, Jun. 2022, https://doi.org/10.2105/AJPH.2022.306831.

[43] C. Caminade et al., "Impact of climate change on global malaria distribution," *Proc. Natl. Acad. Sci.*, vol. 111, no. 9, pp. 3286–3291, Mar. 2014, https://doi.org/10.1073/pnas.1302089111.

[44] M. A. Johansson et al., "An open challenge to advance probabilistic forecasting for dengue epidemics," Nov. 2019, Accessed: Jul. 31, 2026. [Online]. Available: http://hdl.handle.net/10919/105196

[45] L. Hadley, C. Rich, A. Tasker, O. Restif, and S. Funk, "How does policy modelling work in practice? A global analysis on the use of epidemiological modelling in health crises," *PLOS Glob. Public Health*, vol. 5, no. 6, p. e0004675, Jun. 2025, https://doi.org/10.1371/journal.pgph.0004675.

[46] G. Moldovan et al., "AIFS 1.1.0: An update to ECMWF's machine-learned weather forecast model AIFS," Oct. 17, 2025, Atmospheric sciences. https://doi.org/10.5194/egusphere-2025-4716.

[47] A. Meque, S. Gamedze, T. Moitlhobogi, P. Booneeady, S. Samuel, and L. Mpalang, "Numerical weather prediction and climate modelling: Challenges and opportunities for improving climate services delivery in Southern Africa," *Clim. Serv.*, vol. 23, p. 100243, Aug. 2021, https://doi.org/10.1016/j.cliser.2021.100243.

[48] B. Lamptey et al., "Challenges and ways forward for sustainable weather and climate services in Africa," *Nat. Commun.*, vol. 15, no. 1, p. 2664, Mar. 2024, https://doi.org/10.1038/s41467-024-46742-6.

[49] S. Ravuri et al., "Skilful precipitation nowcasting using deep generative models of radar," *Nature*, vol. 597, no. 7878, pp. 672–677, Sep. 2021, https://doi.org/10.1038/s41586-021-03854-z.

[50] Y. Zhang et al., "Skilful nowcasting of extreme precipitation with NowcastNet," *Nature*, vol. 619, no. 7970, pp. 526–532, Jul. 2023, https://doi.org/10.1038/s41586-023-06184-4.

[51] G. Nearing et al., "Global prediction of extreme floods in ungauged watersheds," *Nature*, vol. 627, no. 8004, pp. 559–563, Mar. 2024, https://doi.org/10.1038/s41586-024-07145-1.

[52] R. Pande and M. Jagnani, "Experimental Evaluation of a Flood Forecasting Tool" *AEA Registry*, 2019, https://doi.org/10.1257/rct.4947-9.0.

[53] C. Aitken et al., "Designing probabilistic AI monsoon forecasts to inform agricultural decision-making," Mar. 10, 2026, *arXiv*, https://doi.org/10.48550/arXiv.2603.07893.

[54] Y. Q. Sun, P. Hassanzadeh, M. Zand, A. Chattopadhyay, J. Weare, and D. S. Abbot, "Can AI weather models predict out-of-distribution gray swan tropical cyclones?," *Proc. Natl. Acad. Sci.*, vol. 122, no. 21, p. e2420914122, May 2025, https://doi.org/10.1073/pnas.2420914122.

[55] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke, "Numerical models outperform AI weather forecasts of record-breaking extremes," Aug. 21, 2025, *arXiv*, https://doi.org/10.48550/arXiv.2508.15724.

[56] L. Olivetti, G. Messori, P. Avner, and S. Hallegatte, "Assessing The Real-World Economic Value of Weather Forecasts under Compounding Extremes," *World Bank Work*. Pap., 2026.

[57] Y. Q. Sun, P. Hassanzadeh, T. Shaw, and H. A. Pahlavan, "Predicting Beyond Training Data via Extrapolation versus Translocation: AI Weather Models and Dubai's Unprecedented 2024 Rainfall," Feb. 02, 2026, *arXiv*, https://doi.org/10.48550/arXiv.2505.10241.

[58] M. Mardani et al., "Residual corrective diffusion modeling for km-scale atmospheric downscaling," *Commun. Earth Environ.*, vol. 6, no. 1, p. 124, Feb. 2025, https://doi.org/10.1038/s43247-025-02042-5.

[59] H. Sun et al., "China Regional 3 km Downscaling Based on Residual Corrective Diffusion Model," *EGUsphere*, pp. 1–26, Feb. 2026, https://doi.org/10.5194/egusphere-2026-822.

[60] J. D. L. Brazidec et al., "Downscaling weather forecasts from Low- to High-Resolution with Diffusion Models," Mar. 30, 2026, *arXiv*, https://doi.org/10.48550/arXiv.2604.03303.

[61] R. Molinaro et al., "Universal Diffusion-Based Probabilistic Downscaling," Apr. 20, 2026, *arXiv*, https://doi.org/10.48550/arXiv.2602.11893.

[62] J. Lee and S. Shamekh, "How far can we downscale? Resolution limits and physical interpretability of diffusion models for African precipitation," *Mach. Learn. Earth*, vol. 2, no. 1, p. 015018, Jun. 2026, https://doi.org/10.1088/3049-4753/ae7114.

[63] J. Pathak et al., “Learning Accurate Storm-Scale Evolution from Observations,” Jan. 24, 2026, *arXiv*, https://doi.org/10.48550/arXiv.2601.17268.

[64] A. Allen et al., “End-to-end data-driven weather prediction,” *Nature*, vol. 641, no. 8065, pp. 1172–1179, May 2025, https://doi.org/10.1038/s41586-025-08897-0.

[65] N. Van De Giesen, R. Hut, and J. Selker, “The Trans-African Hydro-Meteorological Observatory (TAHMO),” *WIREs Water*, vol. 1, no. 4, pp. 341–348, Jul. 2014, https://doi.org/10.1002/wat2.1034.

[66] H. Leijnse, R. Uijlenhoet, and J. N. M. Stricker, “Rainfall measurement using radio links from cellular communication networks,” *Water Resour. Res.*, vol. 43, no. 3, p. 2006WR005631, Mar. 2007, https://doi.org/10.1029/2006WR005631.

[67] C. F. Mass and L. E. Madaus, “Surface Pressure Observations from Smartphones: A Potential Revolution for High-Resolution Weather Prediction?,” *Bull. Am. Meteorol.* Soc., vol. 95, no. 9, pp. 1343–1349, Sep. 2014, https://doi.org/10.1175/BAMS-D-13-00188.1.

[68] R. Maulik et al., “Efficient high-dimensional variational data assimilation with machine-learned reduced-order models,” *Geosci. Model Dev.*, vol. 15, no. 8, pp. 3433–3445, May 2022, https://doi.org/10.5194/gmd-15-3433-2022.

[69] ECMWF, “Fifty years of data assimilation at ECMWF.”

[70] C. J. Steele, P. Gill, and M. Spurrier, “Exploring the Use of Public Weather Station Data for Operational Weather Forecast Verification,” *Meteorol. Appl.*, vol. 32, no. 5, p. e70086, Sep. 2025, https://doi.org/10.1002/met.70086.

[71] N. D. Brenowitz et al., “Climate in a Bottle: Towards a Generative Foundation Model for the Kilometer-Scale Global Atmosphere,” Jul. 30, 2025, *arXiv*, https://doi.org/10.48550/arXiv.2505.06474.

[72] M. Alexe et al., “ECMWF Newsletter,” Data-driven ensemble forecasting with the AIFS.

[73] G. Lentze, “ECMWF’s AI forecasts become operational.” Accessed: Jul. 31, 2026. [Online]. Available: https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ai-forecasts-become-operational

[74] T. Rieutord, C. Mignon, E. McAufield, and E. Gleeson, “Energy and carbon footprint considerations for data-driven weather forecasting models,” *Weather*, vol. 81, no. 4, pp. 114–119, Apr. 2026, https://doi.org/10.1002/wea.70035.

[75] D. Patterson et al., “Carbon Emissions and Large Neural Network Training,” Apr. 23, 2021, *arXiv*, https://doi.org/10.48550/arXiv.2104.10350.

[76] “AI’s carbon footprint is bigger than you think | MIT Technology Review.” Accessed: Aug. 03, 2026. [Online]. Available: https://www.technologyreview.com/2023/12/05/1084417/ais-carbon-footprint-is-bigger-than-you-think/

[77] “Why the climate promises of AI sound a lot like carbon offsets | MIT Technology Review.” Accessed: Aug. 03, 2026. [Online]. Available: https://www.technologyreview.com/2025/04/10/1114912/why-the-climate-promises-of-ai-sound-a-lot-like-carbon-offsets/

[78] R. Masiwal et al., “Decision-oriented benchmarking to transform AI weather forecast access: Application to the Indian monsoon,” Feb. 03, 2026, *arXiv*, https://doi.org/10.48550/arXiv.2602.03767.

[79] D. S. Richardson, “Skill and relative economic value of the ECMWF ensemble prediction system,” *Q. J. R. Meteorol.* Soc., vol. 126, no. 563, pp. 649–667, Jan. 2000, https://doi.org/10.1002/qj.49712656313.

[80] T. Dinku, P. Ceccato, E. Grover-Kopec, M. Lemma, S. J. Connor, and C. F. Ropelewski, “Validation of satellite rainfall products over East Africa’s complex topography,” *Int. J. Remote Sens.*, vol. 28, no. 7, pp. 1503–1526, Apr. 2007, https://doi.org/10.1080/01431160600954688.

[81] T. Dinku et al., “ENACTS: Advancing Climate Services Across Africa,” *Front. Clim.*, vol. 3, p. 787683, Jan. 2022, https://doi.org/10.3389/fclim.2021.787683.

[82] E. Coughlan de Perez, J. Arrighi, and J. Marunye, “Challenging the universality of heatwave definitions: gridded temperature discrepancies across climate regions,” *Clim. Change*, vol. 176, no. 12, p. 167, Nov. 2023, https://doi.org/10.1007/s10584-023-03641-x.

[83] T. Dinku et al., “Enhancing National Climate Services (ENACTS) for development in Africa,” *Clim. Dev.*, vol. 10, no. 7, pp. 664–672, Oct. 2018, https://doi.org/10.1080/17565529.2017.1405784.

[84] A. Grossi and T. Dinku, “Enhancing national climate services: How systems thinking can accelerate locally led adaptation,” *One Earth*, vol. 5, no. 1, pp. 74–83, Jan. 2022, https://doi.org/10.1016/j.oneear.2021.12.007.

[85] R. Faniriantsoa and T. Dinku, “ADT: The automatic weather station data tool,” *Front. Clim.*, vol. 4, p. 933543, Aug. 2022, https://doi.org/10.3389/fclim.2022.933543.

[86] T. Dinku, R. Faniriantsoa, S. Islam, G. Nsengiyumva, and A. Grossi, “The Climate Data Tool: Enhancing Climate Services Across Africa,” *Front. Clim.*, vol. 3, p. 787519, Feb. 2022, https://doi.org/10.3389/fclim.2021.787519.

[87] Y. Sophia et al., “Dengue dynamics, predictions, and future increase under changing monsoon climate in India,” *Sci. Rep.*, vol. 15, no. 1, p. 1637, Jan. 2025, https://doi.org/10.1038/s41598-025-85437-w.

[88] L. Cooley and J. Linn, “Scaling Fundamentals: A Framework for Understanding and Managing the Scaling Process,” *Scaling Community of Practice*, May 2024. Accessed: Aug. 08, 2026. [Online]. Available: https://scalingcommunityofpractice.com/scaling-fundamentals/

[89] AIM for Scale, “Agricultural Innovation Mechanism for Scale,” *AIM for Scale*. Accessed: Aug. 08, 2026. [Online]. Available: https://aimforscale.org

[90] A. Jina and P. Winters, “AI is transforming weather forecasting – and that could be a game changer for farmers around the world,” *The Conversation*. Accessed: Aug. 08, 2026. [Online]. Available: https://theconversation.com/ai-is-transforming-weather-forecasting-and-that-could-be-a-game-changer-for-farmers-around-the-world-263030

[91] P. Baker, “An IDA Health Window: The Future of Global Health Financing,” Center for Global Development, Washington, DC, Policy Brief, 2026. Accessed: Aug. 08, 2026. [Online]. Available: https://www.cgdev.org/publication/ida-health-window-future-global-health-financing

[92] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data,* vol. 3, no. 1, p. 160018, Mar. 2016, https://doi.org/10.1038/sdata.2016.18.

[93] International Telecommunication Union, “Statistics,” ITU-D ICT Statistics. Accessed: Jul. 31, 2026. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/pages/stat/default.aspx

[94] African Union, “Continental AI Strategy,” Jul. 2024.

[95] C. Brandon, B. Kratzer, and A. Aggarwal, “Adapting to Climate-Related Health Risks: The Economic Case for Climate Services for Health,” *World Resour. Inst.*, Mar. 2026, https://doi.org/10.46830/wriwp.25.00120.

Appendix: The Opportunities In Detail

Each recommendation from Chapter 7, described in full. Numbering follows the table in Chapter 7. Chapter numbers signify where the gap or opportunity is discussed in the main text. Throughout the process leading to this report, the primary selection criterion was to fund forecasts and systems by whether they change a health decision, not by whether they reproduce the atmosphere.

DISCOVERY

Catalytic capital to steer AI weather toward health.

PUBLIC GOODS AND HEALTH GAINS

RECOMMENDATION R-D1

A health-and-impacts gold-standard dataset.

A reference dataset for health-relevant evaluation, so that decision-oriented benchmarking, evaluation and health modeling have something to anchor on the way reanalysis anchored weather modeling. The direct analog is ECMWF's ERA5, on which all first-generation AI weather models were trained and evaluated. It unlocks the model-selection decision behind every health forecast, which is which forecast to trust for a given outcome and lead time. It has no single owner, so it must be governed as a public good, which raises the open questions of who sets inclusion criteria, how bias and missingness are documented, and how it avoids privileged settings that already have strong surveillance. A better reference corpus lets decisions be judged against health-relevant truth rather than atmospheric proxies. [medium-term] (Ch. 2, Ch. 4)

RECOMMENDATION R-D2

A ground-truth weather reference dataset for data-sparse regions.

Distinct from R-D1 and addressing the other side of the same problem. Reanalysis is weakest in the tropics because there were fewest observations to reconstruct from, so the reference record against which models are scored is thinnest exactly where the climate-sensitive health burden is heaviest. A model can score well there and still be wrong. Assembling an observation-based record for those regions improves the training data and the standard models are judged against at the same time, which makes it one of the few investments that improves the forecast and the evaluation together. It serves the national service's decision about which model to trust locally, and is owned by the services and regional bodies holding the observations. ENACTS demonstrates the method at national scale. [medium-term] (Ch. 2, Ch. 4)

RECOMMENDATION R-D3

Decision-oriented benchmarking of forecasts for health.

A platform and tools that evaluate forecasts against the decision, beginning with heat and extending to other outcomes, and covering both health-relevant thresholds on meteorological variables and, where the data exist, health outcomes directly. It serves the decision of which model a service adopts for a specific health decision and lead time. It is collectively owned, hence a public good and a governance question: who runs the platform, how metrics are validated, and how it stays model-agnostic and vendor-neutral. Making value visible rather than only accuracy is what steers future models and investment toward health. [near-term] (Ch. 3, Ch. 4)

RECOMMENDATION R-D4

A climate-health model intercomparison project (CH-MIP).

The analog to the climate community's CMIP and the crop community's AgMIP, similar in spirit to the CDC West Nile and dengue challenges but established as a recurrent institution and built around multiple health outcomes so that it also determines which models are fit for which timescales and decisions. It serves the decision of which health or epidemiological model to trust for a policy application, by standardizing inputs, outcomes and evaluation. Its own governance is the open design question this report raises: who convenes it, how recurrence is funded, and how modeling groups in the countries carrying the greatest burden shape it rather than respond to it. A recurrent intercomparison builds a modeling community instead of one-off challenges. [medium-term] (Ch. 2, Ch. 4)

RECOMMENDATION R-D5

Shared tooling for adapting and evaluating forecast-improvement techniques.

Adapting a pre-trained model to a place and a purpose, whether by calibration, bias-correction, post-processing, fine-tuning, or weighting the training objective toward the tail, currently proceeds ad hoc and with mixed results. The same tooling is needed twice over, and that is why one investment covers both: a research group comparing which technique delivers benefit for which hazard at what cost needs it to run the comparison, and a national service adapting a model to its own data and decisions needs it to run in production. Building it once serves both, and the results of the first tell the second what to use. It serves the practitioner's decision to act on a forecast at all, and the funder's decision about which improvement is worth its cost. It is owned in use by national services and their health counterparts, and it will not be self-funded by the publishable frontier, which rewards new modeling techniques rather than the operational science that makes them usable. Better plumbing turns a technically skilful forecast into one a district officer can act on. [near-term] (Ch. 3)

FRONTIER CAPABILITY

RECOMMENDATION R-D6

A subseasonal-to-seasonal (S2S) forecasting-for-health consortium.

Pursuing the high-value-if-achievable frontier lead time of two weeks to two months for health decisions. It differs from atmospheric-science S2S work because that work is judged against the very hard benchmark of atmospheric accuracy at these ranges, whereas decision-relevant evaluation may show skill. It serves anticipatory decisions in that window, including pre-positioning, procurement and campaign timing. The owner of those decisions varies by outcome and country, and the consortium's own governance is an open design question. A health-focused agenda will not form organically inside a single research center, and if the frontier lead time becomes usable it opens decisions currently unsupported at any horizon. [medium- to long-term] (Ch. 2, Ch. 3)

RECOMMENDATION R-D7

Training from observations, including satellites, for data-sparse regions.

Forecasting learned directly from satellite radiances, station reports and other observations without conventional data assimilation, with the caveat that tropical performance requires dedicated retraining and has not yet been demonstrated. It serves the short-range warning decision in places with no radar and no operational numerical weather prediction, and is owned by the national and regional services in those regions. A working observation-primary path gives a decision-maker a first usable short-range forecast where none exists today. [medium-term] (Ch. 3)

RECOMMENDATION R-D8

Novel data approaches for AI weather forecasting.

These failed a cost-benefit test under conventional prediction because assimilation was costly and labor-intensive; because assimilation is no longer costly for AI models, the cost falls sharply while the benefit may hold or grow. What remains almost entirely unmeasured is the value of any particular source for a particular forecast, region and lead time. It serves the question upstream of many decisions, which is whether cheap data now add usable skill, and is owned by services deciding where to spend scarce observation budgets. If cheap data prove skilful, they widen who can produce a decision-relevant forecast at all. [medium-term] (Ch. 3)

RECOMMENDATION R-D9

Generative scenario models conditioned on health-relevant compound events.

Models that can be run backwards from a specified outcome to ask what produces it, how likely it is, what conditions precede it, and how far ahead those conditions announce themselves. Heat arriving with poor air quality, or over a district that has just flooded, is the case that matters and the case a record not yet containing the event cannot support. It serves the planning decision that cannot currently be made

at all, and is owned jointly by the modeling groups building the models and the health authorities specifying which compound events matter. Benchmarking on precursor lead time rather than on atmospheric accuracy is what makes this a health investment and not a climate modeling one. [medium- to long-term] (Ch. 3)

RECOMMENDATION R-D10

Regional model-training and localization centers.

AI model training is costly and infrequent, and differs from operational running. Many weather needs span borders, so coordinated effort across a large domain yields more to each country than working alone. Regional bodies may therefore be a better home for retraining and localization than any single national service. It serves the decision about where retraining capacity should sit, and is owned by regional centers and the national services they serve. Where this works, sovereignty is gained without every country carrying the full cost of it. [medium-term] (Ch. 3)

RECOMMENDATION R-D11

An open-source or benchmarkable standard for AI weather models.

A norm rather than a facility. Open-source models, tools and frameworks are a necessary condition for gains in data-limited regions, because they let operational users benchmark, tailor, combine and generate forecasts for their own needs. Where models are closed, a national service cannot verify what it is running or adapt it to what it needs. It serves every downstream evaluation and adaptation decision, and is owned by the model developers, the private sector chief among them, with funders and procuring authorities able to require it. Advocacy and procurement conditions are the instruments here. [near-term] (Ch. 3, Ch. 4)

EVIDENCE

Capital to establish whether forecasts change actions and outcomes, with evidence preceding scaling. This category is public goods throughout: evaluation is the classic under-supplied good, valuable to the whole field and owned by none of it.

RECOMMENDATION R-E1

Impact evaluation of health benefits from forecasts, including assessment of communication and the decision boundary.

Evaluation and feedback built into climate services for health from the start, testing three outcome classes: belief, understanding and trust updating; behavioral change; and health outcomes. This ensures that even with inconsistent data ecosystems and rare health outcomes, forecast effectiveness and uptake can be understood and improved. The communication half matters as much: a forecast that is accurate at the atmosphere can still be damaging at the point of action if it raises alarm without a feasible response, or if repeated false alarms spend the trust a warning system runs on. Uncertainty communication, decision-threshold design and accountability frameworks for action under uncertainty should be developed alongside, beginning with heat, where action thresholds are furthest developed. It serves the decision-maker's act of setting and defending a threshold, and the funder's decision about whether a service is worth continuing. The owner is the body that issues the warning, jointly with the health authority that acts on it. Warning systems are rarely evaluated for whether they change health outcomes, and built-in evaluation turns anecdote into evidence over the first cycles of a service. [near-term] (Ch. 2, Ch. 4)

RECOMMENDATION R-E2

An evaluation and match-making facility.

Instead of relying on scattered researcher proposals, a dedicated stream that matches evaluation researchers with implementers, builds evaluation into programs already being funded as opposed to commissioning it separately afterwards, and targets high-value questions: dengue early-warning impact evaluation, or heat early warning tested for whether a forecast changed behavior, whether employers gave heat breaks, whether productivity held, and whether harm was prevented. Message-design A/B testing, near-absent in weather-health work despite running at national scale in agriculture, should be included; iterative testing of program variants produces causal evidence in weeks or months rather than years and can run during scale-up [29]. It serves the funder's and the researcher's decision about which chains to test, and is owned by a coordinating facility, which prevents it from becoming a set of scattered proposals. It makes rigorous evaluation happen where it otherwise would not, and evaluation built in at rollout costs a fraction of generating the same evidence afterwards. [near- to medium-term] (Ch. 2)

RECOMMENDATION R-E3

Demand articulation as a funded, first-class deliverable, with research on protective behaviors at newly available lead times.

Three things that are one investment. Identifying the decisions a service should be built around, before anything is built. Research on what protective actions people take at lead times that have not previously been available, and on optimal warning timing, both of which the literature review, expert interviews and convening found genuinely empty. And funding the scaling of the method once it works. Demand articulation exercises are inexpensive relative to the national investments they steer, and asking what forecast products are wanted returns only what people already know is available, so the method needs a step that shows people capability they have not seen. It serves the design of every warning downstream, and is owned by the health sector that states the decisions. The WISER case in Chapter 5 suggests a workable model. [medium-term] (Ch. 2, Ch. 5)

SCALING

Development-bank and bilateral capital to roll out what works, unlocked by the economic case.

PUBLIC GOODS AND INSTITUTIONS

RECOMMENDATION R-S1

A climate-health scaling mechanism.

An analog to AIM for Scale, providing design technical assistance to development-bank operations, brokering meteorological-health cooperation, and unlocking large loans with small catalytic grants. It first diagnoses why meteorological-health cooperation has historically stalled, then confirms that health carries budget priority with the target country or the major scaling funders, and aligns government priorities with development-bank funding. It determines whether a loan carries a well-designed meteorological-health component or a copy-pasted one; the decision-owners are governments and bank task teams, and the mechanism brokers between them. Its own governance, ownership and neutrality are open design questions this report raises for co-development rather than a blueprint. With it, small grants unlock far larger loans; without it, proven pilots stall. [medium-term] (Ch. 5, Ch. 6)

RECOMMENDATION R-S2

Unbundled procurement: funding the problem, not the model.

A small problem-definition grants facility followed by a competitive, model-agnostic challenge judged on decision metrics, and not selecting a vendor up front. It serves the procuring authority's decision about what to buy, replacing "pick a vendor" with "define the problem, then run a model-agnostic challenge judged on decision metrics," and is owned by the procuring ministry or agency. Its fundability rests on answering some specific questions: who owns problem definition, how metrics are validated, how vendors are kept out of agenda-setting, and how open standards are enforced. Done right, procurement rewards decision skill as opposed to incumbency. [medium-term] (Ch. 4, Ch. 7)

HEALTH-FACING APPLICATIONS

RECOMMENDATION R-S3

AI weather for early warning folded into existing heat action plans.

Together with a heat-action-plan drafting tool that helps a low-capacity city assemble a plan, feasible with current capacity and data. It serves a city's decision to activate heat measures, including advisories, cooling centers and work-hour rules, and, upstream, a low-capacity city's decision about how to assemble a plan at all. It is owned by a municipal heat-health focal point or disaster-management authority with the authority to trigger those measures and the staffing, facilities and channels to deliver them. A sharper forecast means earlier, better-calibrated activation and fewer wasted or missed triggers. Heat is a good first application, and the transfer path to other outcomes is stated in Chapter 2. [near-term] (Ch. 5)

RECOMMENDATION R-S4

Monitoring and surveillance integration for climate-sensitive disease.

Where a long biological lag sits between the weather and the illness, that lag already supplies the warning time a decision needs, and monitoring current conditions does more work than any forward forecast. Investment has followed the forecast rather than the lead time the decision requires. Integrating environmental monitoring into existing disease surveillance is the cheaper and more immediately useful investment for vector-borne and waterborne disease, and it is also what generates the health data that benchmarking and evaluation later depend on. It serves the control program's timing decision, and is owned by the disease program and the surveillance system it runs on. [medium-term] (Ch. 2)

RECOMMENDATION R-S5

Weather as a use case for digital public infrastructure.

Using weather demand to motivate development banks and governments to build the underlying registries, identity and payment systems, and communication networks that health services depend on, rather than funding those directly. Weather information is a broad-based use case with a wide constituency, which makes it a plausible entry point for infrastructure that is hard to justify on any single sector's business case. It serves the decision to reach affected populations at all, and is owned by the governments and development banks building that infrastructure. Weather is one practical, fundable use case that motivates these rails; it is not the rationale for them, and they must serve wider health and public-interest goals. [medium- to long-term] (Ch. 3, Ch. 6)

SUSTAINABILITY

Workforce, institutions and data infrastructure, on national budgets, not grant cycles. Public goods and institutions throughout: the shared foundations that grant cycles cannot sustain.

RECOMMENDATION R-U1

Capacity realignment.

AI-weather training forums, summer schools and embedded fellowships, so that services benchmark and adapt on their own data and staff feel enabled rather than replaced. The scarce skill is not model development, which will stay concentrated in a small number of centers, but operational use: knowing which model to run for which decision, how to calibrate it locally, how to read what it says, and how to explain that to a health officer. It serves the national service's decision to benchmark and adapt models on its own data so that it doesn't have to depend on outside production, and is owned by the national service or regional body and its host government. It is also the cheapest demand-generation mechanism available, because people who can run these models start asking for things they would not otherwise have known to want. [near-term, ongoing] (Ch. 3, Ch. 6)

RECOMMENDATION R-U2

Health in national AI and data policy frameworks.

A task force to embed health in the AI and data policies being drafted now, beginning with a landscape mapping, and covering three things that belong together. Tiered regulation and accountability, in which verification requirements are set by what a decision puts at stake, with the categories of decision requiring human sign-off defined in advance. Data licensing and benefit-sharing frameworks that protect investment while keeping data available, since clear licensing is what makes the observation and dataset recommendations politically possible. And the six provisions a national policy should require: sovereign data spaces, interoperability against open standards, tiered verification by decision class, red teaming before deployment, multi-channel and multi-language delivery as a procurement condition, and local calibration against national observations. It serves the drafting decision itself, and is owned by the ministries writing AI and data policy jointly with the health ministries rarely present at that table. Policy built on enduring principles outlasts the technologies it governs. [near-term] (Ch. 4, Ch. 6)

RECOMMENDATION R-U3

Health data sharing, including federated methods, and the longer-term data-infra-structure fix.

Two mechanisms inside existing institutions. Strengthen the data-sharing systems that already exist rather than building alongside them: DHIS2 runs across roughly eighty countries and the mandate to share across institutions is already written into how it is meant to work, and what has not happened is the actual sharing. And enable health-data use without moving raw records, by publishing learned relationships as derived products or using federated approaches, which addresses the legal barrier without harmonizing case definitions or eliminating re-identification risk. Sensitivity, not mandate, is usually the binding constraint on health-data sharing, and arrangements that address sensitivity directly move faster than those asserting a right of access. It serves every decision downstream of health data, and is owned by health ministries jointly with the hospitals, insurers and private providers who hold much of the record. [medium- to long-term] (Ch. 2, Ch. 6)

RECOMMENDATION R-U4

An institutional responsibility model and stakeholder map.

Governance of AI-driven climate services for health currently has no map. Which institution should hold each of the nine capacities varies too much by country for a general recommendation, but what does not vary is that each needs a responsible and mandated owner, and that ownership of the action a forecast triggers, and of the decision not to act, is rarely assigned to anyone. A reference model that a country can adapt, naming the capacities, the plausible institutional homes for each, and the arrangements that have to be settled between them, is a public good that no single country will produce for itself. It serves the decision about who is answerable when a forecast is wrong or a correct forecast is ignored, and settling that deliberately is better than settling it by the first litigated case. [near-term] (Ch. 6)

The Institute for Climate and Sustainable Growth leverages the University's unique legacy and resources to balance the risks of a changing climate with the essential need for human progress. It does so by combining frontier research in economics and policy, climate systems engineering, and energy technologies with a pioneering approach to energy and climate education. The Institute also seeds interdisciplinary research that explores new topics in this ever-evolving field and deploys practical, effective solutions in countries central to this challenge.

THE UNIVERSITY OF CHICAGO
INSTITUTE FOR CLIMATE & SUSTAINABLE GROWTH

The University of Chicago
5801 S. Ellis Ave.
Chicago, IL 60637
climate.uchicago.edu

Supported by The Rockefeller Foundation